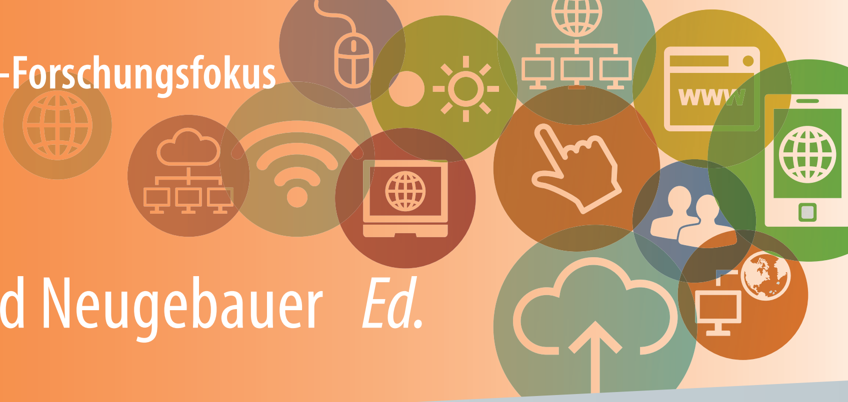
Fraunhofer-Forschungsfokus



Reimund Neugebauer *Ed.*

Digital Transformation

Springer Vieweg

Digital Transformation

Fraunhofer-Forschungsfokus:

Reimund Neugebauer

Digital Transformation

1st ed. 2019

Springer Vieweg

Reimund Neugebauer
Zentrale der Fraunhofer-Gesellschaft,
Munich, Germany

ISBN 978-3-662-58133-9 ISBN 978-3-662-58134-6 (eBook)
https://doi.org/10.1007/978-3-662-58134-6

Springer Vieweg

Printed on acid-free paper

This Springer Vieweg imprint is published by Springer-Verlag GmbH Germany and is a part
of Springer Nature. The registered company is Springer-Verlag GmbH Germany, Heidelberger
Platz 3, 14197 Berlin, Germany

Contents

1 Digital Information – The “Genetic Code” of Modern Technology

Prof. Reimund Neugebauer
President of the Fraunhofer-Gesellschaft

1.1 Introduction: Digitization, a powerful force for change

The digital era began relatively slowly. The first programmable computer using binary code was the Zuse Z3, designed and built by Konrad Zuse and Helmut Schreyer in Berlin in 1941. In 1971 the first microprocessor was patented; it contained 8,000 transistors. Within ten years, nearly ten times as many transistors were being used; by 2016, the number was around 8 billion.

This exponential growth in complexity and power of digital computers was predicted by Gordon Moore in 1965. *Moore's Law*, a rule of thumb to which he gave his name, is generally held to mean that the number of transistors that fit into an integrated circuit of a specified size will double approximately every 18 months. While the law's interpretation and durability may be the subject of debate, it nevertheless provides a fairly accurate description of the dynamics of the development of digital technology.

Among the reasons for this vast acceleration in progress is the fact that digitization, as well as changing various technical and practical fields of application, has also changed the work of research and development itself. A processor featuring transistors running in the billions, for example, can only be developed and manufactured using highly automated and digitized processes. Complex computer programs are themselves in turn designed, realized and tested in whole or in part by computers. The immense quantities of data generated by research projects, production facilities and social media can only be analyzed with massive computer assis-

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_1

tance. The insights we gain from this analysis, however, were practically unattainable just a few years ago. Now, Machine Learning is becoming standard: Artificial systems gather experiences and are then able to generalize from them. They produce knowledge.

But the development momentum is also reinforced by the fact that technological fields of application are expanding equally quickly. The need for digital systems appears inexhaustible, because they contribute to improvements in nearly every area regarding performance, quality and (resource) efficiency of products and processes. The developmental leap is so all-encompassing that we can justifiably speak of a "digital revolution".

1.2 Technology's "genetic code"

Machines need instructions in order to function. For simple processes this can be achieved by manual operation, but this no longer meets the demands of modern production machinery and facilities. Numerous sensors provide vast quantities of data which the machine stores and interprets and responds according to rules stored in a digital code. Such systems add up the knowledge of their developers, the results of Machine Learning and current data.

Biological organisms, too, gather numerous data from their environment and interpret it. The blueprint for their construction – the genetic code found in each and every cell – represents the aggregated knowledge gathered over the whole course of the organism's evolution.

Ideally, for both organisms and machines, the information collected and stored holds the answers to every conceivable challenge. For this reason, and in spite of numerous differences at the level of detail, the digital code's function and operation is reminiscent of that of the genetic code of biological systems. In each case:

- The code contains the information for the actions, reactions and characteristics of the organisational unit, whether it is a cell or a machine.
- Complex information is compressed into a comprehensive sequence of a limited number of 'characters'. DNA manages with four characters, the nucleotides adenine, guanine, thymine and cytosine; the digital code uses just two, the numbers 0 and 1.
- These codes can be used to store not only the construction and behavioural frameworks of smaller units, but of overall structures, too – the structures of entire organisms in the case of the genetic code, and of a production plant or factory in the case of the digital code. Both systems have at their core the capacity for flexibility and learning.

- The stored information can be copied limitless and practically lossless. Identical DNA replication is done by splitting of the double helix and completing the free binding sites by attaching new complementary DNA-bases.
- Digital information copying happens through lossless reading and renewed recording of the information on a storage medium.
- The information is retained during duplication. It may however be modified if there is a need for adaptation: In the case of the genetic code, this takes place via mutation, for example, or through the recombination of existing partial information with subsequent selection; in the case of the digital code, parts of the scripts can be replaced or expanded.

In this way, both the digital and genetic codes carry the potential for conservation as well as innovation, and both are combinable – changes can thus build on already existing achievements. This is the explanation why digitization has led to such a boost in innovation. It leverages the potential of evolutionary progress through research and development onto an unprecedented level in the field of technology.

What reinforces the effectiveness of the digital code as a driver of evolution and innovation even further – in comparison to the genetic code – is the fact that on the one hand new digital information can be integrated in a very targeted manner and, on the other hand, be transported worldwide in real time via the Internet. Evolutionary improvements of technologies can thus happen quickly and be made available immediately, hindered only by restrictions of patent law and politics.

1.3 The dynamics of everyday digital life

As a consequence, digitization has a tremendous dynamizing effect on the current and future development of technology. It is only ten years since the first smartphone came onto the market, but people's lives have already changed enormously across the globe as a result of this single invention. No matter where we are, we can connect visually and audibly with whoever we want. The impact on our home and work lives – and even mobility and migration behaviors – is already clearly visible.

The whole working and living environment is in a state of upheaval. Highly sophisticated control units are increasing the efficiency, speed and performance of practically every technological device we are using each day. Mobility, energy-efficient air-conditioning, automated household appliances, the ubiquitous availability of communication and working opportunities, of information and entertainment, to name just a few, provide us with previously unimaginable opportunities. The development of the tactile Internet enables us to cause some effect in real time on the other end of the world upon the click of a button. Efficient and flexible production technologies allow for individual product design and the manufacturing of many products via personal 3D printers.

The decentralized production of content or goods, as mentioned above, is a remarkable side-effect of digitization. The fast publication of individually created books, images, films, music, objects, ideas and opinions – often without control and censorship – has become commonplace through the Internet. This creates new and rapid opportunities for commercialization and self-actualization, but also leads to societal dangers that we still must learn to deal with.

These new possibilities also generate expectations and demands: We, for example, not only get accustomed to the convenient status quo, but also to the dynamics of development. Tomorrow we expect even more from what today's digitally-based products and media already offer. And that means that the international markets for technological products and the technologies themselves are subject to the same massively increased dynamics of change and growth.

This is also the reason why digitization, far from having a depressive effect on the job market, is instead a driver of it. In contrast to the commonly-held fear that digital technology is destroying jobs, it has first and foremost led to changes in regard to job profiles, and even led to an overall increase in job- and income opportunities. Current and future employees are expected to be more flexible, have the willingness to learn and possibly also show a certain degree of professional curiosity. Companies have to respond faster and more flexibly to market changes, both with new products but also with new, disruptive business models. They need to anticipate the future requirements of the market so they can offer the right products at the right time.

Lifelong learning has become an indispensable reality for everyone who participates in the economic process, and digitization – along with the associated dynamization of developments – is the principal cause.

1.4 Resilience and security

Because of digital technology, nearly all areas that are essential for business, science, public or private life are nowadays controlled technologically: security, healthcare, energy supply, production, mobility, communications and media. The more areas we leave to the control of data technology, the greater the importance for its reliability. This applies to individual systems such as cars and airplanes as well as to complex structures like supply systems and communications networks. With the advance of digitization, resilience, i.e. the ability of a system to continue functioning even in the case of failure of individual components – is thus becoming a key development goal.

Today, information is stored and transported almost exclusively in digital form. But data is not just something we consume, it is also something we produce, like every digitally controlled product and production facility is doing. The volume of digital data produced every day is constantly increasing. This data is both informative and useful – and thus valuable. This applies to personal data as well as data produced by machines, which can be used to explain, improve or manage processes. This is also the reason why digital data is a trading commodity making it already one of the most valuable goods of the 21st century.

Automated driving – a specific technological vision that can only be realized with the aid of highly advanced digital technology – can only truly become a reality when it is possible to unhesitatingly and permanently hand over complete control to the car. This requires a huge amount of automated communication to take place reliably and seamlessly, i.e. between the steering control and sensors of the car, between traffic participants, and also with infrastructure such as traffic management systems and location services.

And that is only one example of how much modern products and infrastructure are depending on the proper functioning of digital technology. This applies to the entire information- and communication technology, manufacturing and health technologies as well as logistics and networked security systems. Therefore, it is no exaggeration to say that digital technology has already become the foundation of a technologically-oriented civilization.

This leads to the conclusion, that safety and security are the key issues when it comes to digitization. Designing products, systems and infrastructures in such a way that they will work constantly and without exception in the interest of humans will become a central goal for technological development.

It is here that the Fraunhofer-Gesellschaft – with its wide-ranging competences in the fields of information technology and microelectronics – sees a key challenge and an important work area for applied research.

In the area of digitization, the terms ‘safety’ and ‘security’ are differentiated and refer to *operational safety* and *security from attack*. In both areas a permanently high research need is required. Experience shows that cyberattacks on digital infrastructure utilize latest technologies, which means that safety and security need to be reinforced through continuous research to be always one step ahead of the attackers. In view of the potential harm from modern cyberattacks and large-scale IT infrastructure outages, a relatively significant investment in safety and security research seems warranted.

Likewise, it is also important to support (further) training of experts and to create advanced security consciousness among the relevant specialists and professional users. The Cybersecurity Training Lab – a concept rolled out in nine locations by Fraunhofer, in partnership with universities, and sponsored by the German Federal Ministry of Education and Research (*Bundesministerium für Bildung und Forschung*, BMBF) – is one example of such an endeavor.

Politics and society, too, need to be better sensitized and informed with regard to data protection. Last but not least, private individuals are called upon to make digital security as much second nature as locking doors and windows. Technologies that in many ways encourage decentralization also demand greater responsibility on the part of individuals, in regard to security as well as ethics.

1.5 Fraunhofer searches for practical applications

The Fraunhofer-Gesellschaft has taken on a unique role among scientific organizations. On the one hand it is conducting research with the commitment for scientific excellence, and on the other hand, it is the stated goal to achieve results for practical applications. For this reason, Fraunhofer stands at the forefront when it comes to the invention and development of new technologies. We are pivotal figures in key technologies and make large contributions to further their progress and their dissemination. This entails a particular responsibility since, in modern industrial societies, technologies exercise a decisive impact on human life.

In the field of digitization, Fraunhofer, in a leading role, is involved in initiatives, developments and partnerships of crucial importance. These include the following, sponsored by the Federal Ministry for Education and Research (BMBF):

- The Cybersecurity Training Lab – Fraunhofer is responsible for concept and implementation.
- The Industrial Data Space initiative – aims to facilitate the secure, independent exchange of data between organizations as a precondition for offering smart services and innovative business models. In the meantime, Fraunhofer has added further elements through the development and inclusion of aspects such as the Material Data Space and Medical Data Space.
- The Internet Institute for the Networked Society – Fraunhofer is a member through the Fraunhofer Institute for Open Communication Systems (FOKUS).
- The Research Fab Microelectronics Germany – originating from a concept developed by the Fraunhofer Group for Microelectronics.

With our experience as a market-oriented provider of research and development services, we are defining fields of technology with major current and future significance. They are thus moving into the focus of our activities. We have selected three key research fields that have the potential to impact people’s lives extensively in the future. These are:

- Resource efficiency
- Digitization
- Biological transformation

We have created the Fraunhofer Research Focus series together with Springer Vieweg in order to underline the importance of these three topics and spread awareness of them within science, industry and among the public. The book in your hands right now is the second volume in this series. It gives an overview of key projects at the Fraunhofer institutes in the field of digitization.

Digitization – Areas of Application and Research Objectives

2

Dr. Sophie Hippmann · Dr. Raoul Klingner · Dr. Miriam Leis
Fraunhofer-Gesellschaft

2.1 Introduction

For most readers of this book, digitization has already become a natural part of their everyday life. In essence, "digitization" refers to the binary representation of texts, images, sounds, films and the properties of physical objects in the form of consecutive sequences of 1s and 0s. These sequences can be processed by modern computers at exceptionally high speeds – billions of commands per second.

Digitization acts as a kind of "universal translator", making data from various sources processable for the computer and thus offering a range of possibilities that would otherwise be unthinkable. These include, for example, carrying out complex analyses and simulations of objects, machines, processes, systems, and even human bodies and organs – as is the case with the 3D body reconstruction technology realized by the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, HHI. Digitized data from sensor captured brain signals can even be used to control computers and robots. The inverse is now also possible, using digital signals to produce haptic sensations in prosthetics. Here, digitization acts as a direct link between the biological and cyber-physical worlds.

Although digitization can be used in nearly all fields, this introductory chapter focuses on the following areas of application, with especially significant impacts on people's lives:

- Data analysis and data transfer
- Work and production
- Security and supply

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_2

2.2 Data analysis and data transfer

2.2.1 The digitization of the material world

In order to reconstruct material objects, it is essential to have the information about their composition, structure and form. These parameters can be digitized and reconstructed in computers. In this way, complex machines, materials and pharmaceuticals can be designed and verified for suitability in simulations even before they have been produced as real objects. Virtual reality projections can visualize objects in a detailed way and respond to user interaction. Digitization is thus also gaining the interest of artists and persons engaged in the cultural sector. With the aid of 3D scanning and digitization technologies, valuable artistic and cultural artifacts can be “informationally” retained in detailed, digitized form. This is accomplished at the Fraunhofer Institute for Computer Graphics Research IGD. In this way, cultural artifacts can be made accessible to anyone – also for scientific research – without the risk of damaging the original.

2.2.2 Intelligent data analysis and simulation for better medicine

Also in medicine the digitization of data – e.g. of medical images, textual information or molecular configurations – plays a considerable role. Effective, safe and efficient processes for analyzing large datasets play an important role here. The analysis of multimodal data – i.e. computer-aided comparison of different images (x-rays, MRIs, PETs, etc.), in combination with lab results, individual physical parameters and information from specialist literature – can pave the way for more accurate diagnoses and personalized treatments that cannot be achieved by any specialist alone to this extent. Digitization in medicine is a specialist field of the Fraunhofer Institute for Medical Image Computing MEVIS. Here, as in other fields, big data needs to be transformed into smart data. Likewise, digital assistants are expected to gain the ability to automatically recognize supplementary, missing or contradictory information, and to close information gaps through targeted enquiries.

2.2.3 Maintaining quality at smaller sizes via data compression

The immense spread of digitization has led to an exponential growth in data volume so that despite capacity increases, data transfer bottlenecks still occur. Some projections predict that the volume of data produced globally could multiply tenfold by 2025, in comparison to 2016, reaching as much as 163 zettabytes (a 163 followed by twenty-one zeros, equal to forty-one thousand billion DVDs).

Today, music and video streaming make up a large part of global data transfers. Lossless data compression thus has an important role to play in reducing the size of digital data packages, thereby shortening transfer times and requiring less memory. Since the groundbreaking development of MP3 coding based on research by the Fraunhofer Institute for Integrated Circuits IIS and Fraunhofer Institute for Digital Media Technology IDMT, Fraunhofer has been constantly continuing to further develop audio and video data compression processes, in order to achieve ever better transmission quality with the lowest possible data volumes. Uncompressed audio files, for example, are up to ten times larger than MP3 files with identical sound quality.

2.2.4 Digital radio – better radio reception for everyone

The Fraunhofer IIS also played a key joint role in another accomplishment in the field of audio technologies – digital radio – having developed its basic, broadcasting and reception technologies. Digital radio offers huge advantages over the analog process. Terrestrial wireless reception can be received even without Internet connection and is free of charge. Energy-efficient transmission, interference-free reception, high sound quality, the option of additional data services and extra space for more broadcasters as well, are among its clear benefits. Digital radio also allows for real time communication of traffic and disaster alerts, with high reliability and reach, even when no Internet connection is available. In Europe, digital radio will largely replace analog systems over the coming years. Even in emerging economies such as India the transition is in full swing.

2.2.5 Transferring more data in less time: 5G, edge computing, etc.

Fast data transfer with minimal delays or latency, often referred to as the “tactile Internet”, is the basis for a range of new technological applications. These include: connected machines; autonomous vehicles and objects that can communicate with people and each other in real time; augmented reality applications that feed in up-to-the-minute updates; or specialists who are able to carry out highly complex surgery safely from the other side of the world via telerobots. In the future, global data usage will shift from video and audio data to (sensor) data from industry, mobility and the “Internet of things”. The new 5G cellular communications standard – with Fraunhofer playing a key role in promoting its development, testing and distribution – promises data transfer rates of 10 Gbp per second, ten times faster than today’s 4G/LTE. The requirements for the industrial Internet are high, demanding above all scalability, real-time capabilities, interoperability and data security.

Work on developing and testing new technologies for the tactile Internet is ongoing at the four Berlin-based transfer centers – the Internet of Things Lab (Fraunhofer Institute for Open Communication Systems FOKUS), the Cyber Physical Systems Hardware Lab (Fraunhofer Institute for Reliability and Microintegration IZM), the Industry 4.0 Lab (Fraunhofer Institute for Production Systems and Design Technology IPK) and the 5G Testbed (Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, HHI). In addition to 5G technologies, cloud computing and edge computing also play a key role here. With the latter, a large part of computing is performed within the individual machines and sensor nodes, thus reducing latency, since not all data has to be processed in the cloud.

2.3 Work and production

2.3.1 The digitization of the workplace

Digitization has fundamentally changed our working world and will continue to do so. Today, email (or chat) has nearly completely replaced the classical letter for day-to-day written communications. Nowadays engineers design prototypes on the computer instead of the drawing board, and in future robots and interactive digital assistance systems will stand by our sides to help with everyday tasks. Technologies for recognizing speech, gestures and emotions enable people to communicate with machines intuitively. Discoveries from neuroscience are also helping to identify what people are focusing on when they use machines, thus helping with the development of better, safer and more user-friendly designs and interfaces. The Fraunhofer Institute for Industrial Engineering IAO, for example, is analyzing what happens in the brains of users of technological devices in order to be able to optimize interaction interfaces. In the workplace of the future, interactive cooperation between people and machines will develop further, while humans nevertheless will increasingly take center stage. This will not only change production work but will also enable new processes and services; research in this direction is ongoing at the Fraunhofer IAO’s Future Work Lab.

2.3.2 Digital and connected manufacturing

By now, modern machines have become cyber-physical systems (CPS) – a combination of mechanical, electronic and digital components able to communicate via the Internet. They are fundamental to Industry 4.0: here, manufacturing facilities and systems are constantly connected. Computers, Internet connections, real-time sensor readings, digital assistance systems and cooperative robot systems make up the components of the future production facilities. The Fraunhofer Institute for Machine Tools and Forming Technology IWU is researching and developing innovations for the digital factory.

Digital copies of machines, "digital twins", are also being used. They exist in the virtual space, but are equipped with all characteristics that could be relevant for their operation under real-world conditions. In this way, possibilities for optimization as well as potential errors can be identified early on, and the behavior under changing conditions can be thoroughly tested in advance. Industry 4.0 should contribute to the optimization of processes for resource efficiency, the improvement of working conditions for people, and the affordable manufacturing of customized products.

2.3.3 Turning data into matter

The connection between the digital and physical worlds becomes especially apparent with additive manufacturing, also known as "3D printing". This allows for the transfer of information about the configuration of objects via computer, so that the object can be "materialized" in another location. Similar to the way how living creatures obtain their physical form from the information stored in their DNA, additive manufacturing automatically instantiates objects physically on the basis of digital data.

The diversity and quality of materials and the resolution are constantly improving in additive manufacturing. This process facilitates a relatively inexpensive production of prototypes, spare parts, customized products or tailored personal prosthetics. Benefits of additive manufacturing processes also include efficient use of materials (since the object is constructed directly from a mass of material rather than being "carved" from an existing piece by removing material), and the ability to produce very small or complex geometries. Additive manufacturing processes are thus going to be firmly integrated into Industry 4.0 concepts. The Fraunhofer Institute for Electronic Nano Systems ENAS, together with other Fraunhofer institutes, is addressing these challenges in a major large-scale project.

2.3.4 Cognitive machines are standing by our sides

Artificial intelligence is becoming increasingly multifaceted and is developing towards “cognitive machines” that have the ability to interact, remember, understand context, adapt and learn. Machine Learning has now become a key technology for the creation of cognitive machines. Instead of programming all the steps for solving a problem in advance, the machine is presented with a very large amount of data from which it can automatically recognize patterns and derive principles, and thus improve its performance. This requires fast processors and large data volumes, “big data”. Herewith machines can learn to process natural language, identify the tiniest irregularities in processes, control complex facilities, or discover subtle abnormalities in medical images.

The fields of application for cognitive machines are ubiquitous, ranging from autonomous driving and medical technology to condition monitoring of industrial facilities and electricity generation plants. Fraunhofer institutes are conducting research to improve cognitive machines, focusing on various areas. These include, for example, effective machine learning with small data sets; improving transparency, especially in the case of learning in deep neural networks (“deep learning”); or incorporating physical data and expert knowledge in “grey box” models.

2.4 Security and resilience

2.4.1 Data – the elixir of the modern world

Data is the DNA and fuel of our modern world. As in the example of additive manufacturing, information about the structure and composition of matter gives the object its function and thus its value. Anyone with the data for a 3D model could in principle produce it anywhere. Similar is the case of pharmaceuticals: they are made of common atoms, but their structure – the arrangement of the atoms – is key to their efficacy. Still, the correct configuration and mode of synthesis can take years to identify. Anyone who possesses and controls relevant data and information holds a competitive advantage. The Fraunhofer Big Data and Artificial Intelligence Alliance helps to unearth such data treasures, without losing sight of quality and data protection issues.

The more complex a technological system is, the more susceptible it is to malfunction and the ensuing effects. This is why complex technological systems have to be tuned for resilience; that is, they need to be able to resist malfunctions and, in the case of damage, still be able to function with a sufficient degree of reliability.

The Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI is dedicated to improving the security of our high-tech systems and the infrastructure that depends on them.

2.4.2 Industrial Data Space – retaining data sovereignty

Data security is an extremely important issue. We need to exchange information to carry out research or offer customized services, but data also needs to be protected from unauthorized access. The problem is not only data theft, but also data forgery. The more processes become data-driven, the more serious the effects of erroneous or forged data can be.

The Fraunhofer-Gesellschaft's Industrial Data Space initiative aims to create a secure data space, enabling businesses of different sectors and of all sizes to manage their data assets while maintaining the sovereignty over their data. In order to facilitate the secure processing and exchange of data, mechanisms for protection, governance, cooperation and control are at the core of the Industrial Data Space. Its reference architecture model is intended to provide the blueprint for a range of applications where such secure and controlled data exchange is essential. These applications include Machine Learning; improvements in resource efficiency within manufacturing; road traffic safety; better medical diagnoses; intelligent energy supply management; and the development of new business models and improved public services. Digitization is also important for the transition to sustainable energy provision since the analysis and intelligent management of consumption, availability and load is one of the tasks of digital systems. These issues are one of the key fields of activity of the Fraunhofer Institute for Experimental Software Engineering IESE.

2.4.3 Data origin authentication and counterfeit protection in the digital world

An important topic within digitization, alongside data encryption, is the validation and secure documentation of digital transactions. Data is easier to manipulate than physical objects since it can be quickly copied, and a few lines of computer code can be enough to alter the functioning of an entire system. Forging a paper banknote or a material object on the other hand is far more difficult. The possibilities offered by blockchain technology for authenticating digital entries and transactions are currently being tested and developed.

Blockchain technology became popular through its use in cryptocurrencies (encrypted, digital currencies) such as bitcoin and Ethereum. In essence, the blockchain is a database – similar to the ledger in double-entry accounting – where the series of transactions is recorded chronologically and protected from manipulation by encryption processes. These blockchain databases may also be operated in a decentralized manner, distributed between several different users. The Fraunhofer Institute for Applied Information Technology FIT is currently investigating its potential and developing innovations for blockchain applications.

2.4.4 Cybersecurity as the foundation for modern societies

IT-, data- and cybersecurity are essential to the functioning of digital societies. Key issues include the prevention of unauthorized access to and manipulation of data and data infrastructures as well as securing personal and person-specific data. When it comes to cars (which may now contain more lines of code than an airplane), energy and water provision (managed by computers), highly-networked Industry 4.0 facilities or smart homes, cyber threats can cause serious problems and thus need to be identified and defended against early on.

Digital system outages can lead to the breakdown of entire supply networks, affecting power, mobility, water or food distribution, among others. The Fraunhofer Institutes for Secure Information Technology SIT and for Applied and Integrated Technology AISEC do not only provide technological innovations for the early detection of potential cyber threats with the aid of Machine Learning, but also cybersecurity courses and training labs.

2.4.5 Cybersecurity technology adapted to people

For the sake of security is it important that IT- and cybersecurity applications are designed to be user-friendly and easy to use. If their usage is too complex and laborious then they will not be used at all, thus increasing the risks. This is why the Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE is investigating ways to maximize user-friendliness of information technology and cybersecurity systems and how they can be designed to be as ergonomic as possible. The new “Usable Security” research project aims at extending the current limits of computer system usability. Technology should be people-centered, not the other way around, as has often been the case so far. Only when people are in the focus of interest, a maximum level of actionability and security can be achieved in cyberspace.

2.4.6 People-centered digitization

Pulling back from digitization is unthinkable for a high-tech society, and would likely be tantamount to a catastrophe. But there are still a lot of new applications ahead of us: automated driving, cooperative robots and assistance systems, telemedicine, virtual reality, and digital public services. The Fraunhofer Institutes for Material Flow and Logistics IML and for Transportation and Infrastructure Systems IVI are developing driver assistance systems for safe and reliable automated driving, in the fields of road traffic, agriculture and logistics.

The continued progress of the digital world also poses challenges like the protection of digital data and infrastructures; efficient, the effective and intelligent management of big data; faster data transfer and reduction of latency as well as the further development of processor technologies and computation methods.

The next developmental phase may be characterized by the linking of digital and biological concepts, since genetic and binary codes are similar. Learning-enabled robotic systems, swarm intelligence in logistics, biosensors, 3D printing, and programmable materials already all point towards this direction. The Fraunhofer-Gesellschaft is dedicated to innovation and solutions for challenges in order to improve and drive forward the process of digitization, with humans always remaining at the center.

Virtual Reality in Media and Technology

3

Digitization of cultural artifacts and industrial production processes

Prof. Dieter W. Fellner
Fraunhofer Institute for Computer Graphics Research IGD

Summary

Virtual and Augmented Reality technologies have by now become established in numerous engineering areas of application. Also in the cultural and media fields interactive three-dimensional content is being increasingly made available for information purposes, and used in scientific research. On the one hand, this development is accelerated by current advances in smartphones, tablets and head-mounted displays. These support complex 3D applications in mobile application scenarios, and enable us to capture our real physical environment using multimodal sensors in order to correlate it with the digital 3D world. On the other hand, new automated digitization technologies such as CultLab3D of the Fraunhofer Institute for Computer Graphics Research IGD allow the production of the necessary digital replicas of real objects, quickly, economically and of high quality.

3.1 Introduction: Digitizing of real objects using the example of cultural artifacts

To allow the best possible retention and documentation of our shared cultural heritage, digital strategies were formally established at the political level worldwide. New initiatives such as the iDigBio infrastructure or Thematic Collections Networks in the USA promote the advanced digitization of biological collections. EU

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_3

member states have also been called on by the European Commission to bolster their digitization efforts. This executive priority is part of the Digital Agenda for Europe and underlines the necessity of facilitating large-scale improvements of the online accessibility of historical cultural artifacts [1]. The Digital Agenda identifies the long-term retention of our cultural heritage as one of the key initiatives of the Europe 2020 strategy, with the goal to provide improved access to culture and knowledge via improved usage of information and communications technologies [2]. These initiatives are closely tied to Article 3.3 of the European Union Treaty of Lisbon [3] which guarantees that "Europe's cultural heritage is safeguarded and enhanced".

Despite the legal framework conditions that recognize its significance for society, our cultural heritage is at risk of all sorts of dangers. Recently, a range of natural and man-made catastrophes have highlighted just how fragile our cultural heritage is. Events such as the deliberate destruction of the ancient Semitic city of Palmyra in Syria or of archeological finds in the museum of Mosul, Iraq underscore the necessity for new and faster methods of documentation, leading to a reappraisal of high-definition facsimiles. Moreover, the fact that only a small proportion of all artifacts in exhibition collections is publicly accessible provides further motivation for improving access to information about our cultural heritage [4]. Innovative documentation methods for cultural heritage items are thus gaining ever-increasing significance. This arises, on the one hand, from the desire to enable improved access to unique items, to make collections more easily accessible for research purposes or to a wider audience for example; and on the other hand, from the latent threat of losing them forever through catastrophes or other environmental factors.

Against this backdrop, and in times of digital transformation, the use of 3D technologies in the cultural sphere is becoming increasingly important. This is because they offer a so far unexploited potential use, whether for documentation or retention purposes, for innovative applications in fields as diverse as education and tourism, for optimal accessibility and visualization of artifacts, or as a basis for research and conservation. They also enable physical copies to be produced as a result of highly precise 3D models. The increasing demand for information and communications technologies demonstrates the growing need for research across the whole value chain, from digitization and web-based visualization through to 3D printing.

This research produces tools that stimulate the development of new technologies for digitally processing and visualizing collections, enabling the safeguarding of cultural heritage. Today, the digital capture of two-dimensional cultural treasures such as books, paintings or "digitally-born" collections such as films, photos and audio recordings is already widespread. Examples of extensive efforts towards mass digitization include initiatives such as the Google Books Library Project, which focuses on the scanning of millions of books worldwide, or the Europeana virtual library, which –in keeping with its 2015 goal –already features more than 30 million digitized artifacts. These set the standard for digital access for the end user.

Thus far, digitization activities have nevertheless been mainly restricted to two-dimensional artifacts. Commercially available technologies for the efficient and highly precise digitization of millions of three-dimensional objects such as sculptures or busts are currently lacking. Initiatives here focus primarily on prestigious individual cases – such as the 3D imaging of the world-renowned Nefertiti bust by TrigonArt GmbH in 2008 and 2011 – instead of on entire series of objects. The reason for this is the significant time and cost overhead still required to capture the entire surfaces of objects, including undercuts. According to studies, the time overhead for repositioning the acquisition device, for example, currently lies at around 85% of the entire acquisition time, regardless of the technology used (structured light or laser scanners. In addition, the technological possibilities for capturing specific materials are still restricted.

3.1.1 Automating the 3D digitization process with CultLab3D

In order to make collection items accessible to various user groups in 3D, too, and meet the need of easy-to-use, fast and thus economical 3D digitization approaches, the Fraunhofer Institute for Computer Graphics Research IGD is developing the CultLab3D modular digitization pipeline [5]. This enables three-dimensional objects to be digitized in 3D with micrometer accuracy, via a completely automated scanning process. This is the first solution of its kind to take account of the aspect of mass digitization. The intention is to increase speeds in order to reduce the cost of 3D scans by between ten- to twentyfold. The project is also striving for true to original reproduction at high quality, including geometry, texture and the optical properties of the material.

The modular scanning pipeline (see Fig. 3.1) currently consists of two scanning stations, the CultArc3D and CultArm3D. The digitization process is fully automated, using industrial materials handling technology and autonomous robots to convey items to the corresponding optical scanner technologies. By decoupling the color-calibrated capture of an object's geometry and texture from its final 3D reconstruction via photogrammetry, the scanning pipeline achieves a throughput of just 5 minutes per object. Additionally, with sufficient computing power, a final 3D model can be produced every five minutes. In most cases, little or no post-process-

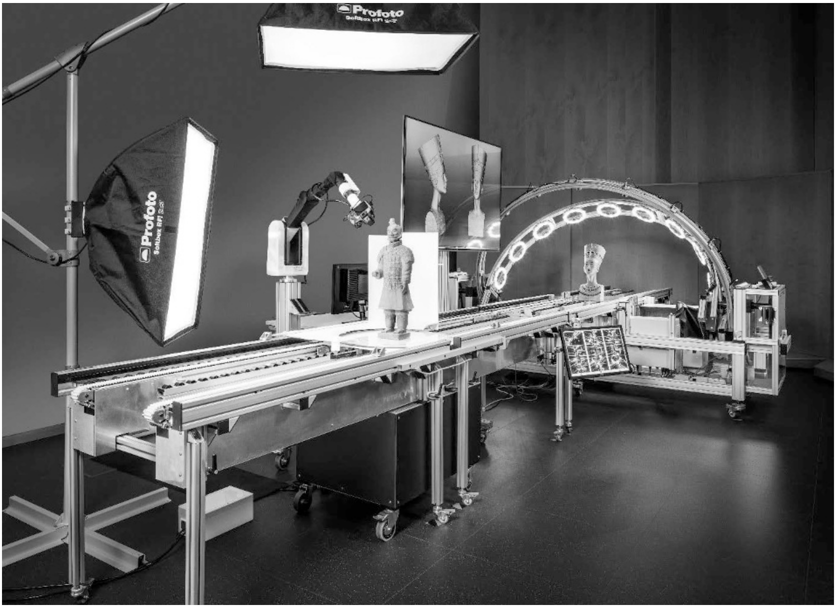


Fig. 3.1 Fraunhofer IGD CultLab3D scanning pipeline (Fraunhofer IGD)

ing is required. The scanning pipeline currently digitizes items of 10–60cm in height or diameter. The entire scanning pipeline is operated by a tablet PC, which all of the system's components log in to.

CultArc3D

The CultArc3D can either be operated alone or in conjunction with other scanners. The module captures both geometry and texture as well as optical material properties according to previous works [6][7]. During the digitization process, a conveyor belt fully automatically moves the objects to be scanned through the CultArc3D, on glass tablets.

The CultArc3D (see Fig. 3.2) consists of two coaxial semicircular arches. These turn around a common axis. Both arches cover a hemisphere that is centered on an object placed in the midpoint. Each arch is moved by its own actuator so that a discrete number of stop positions is possible. The arches have different radiuses so they can be moved independently. The outer arch (hereinafter referred to as the "camera arch") holds nine equidistant cameras. Nine additional cameras beneath the object conveyor surface capture the artifact's underside.

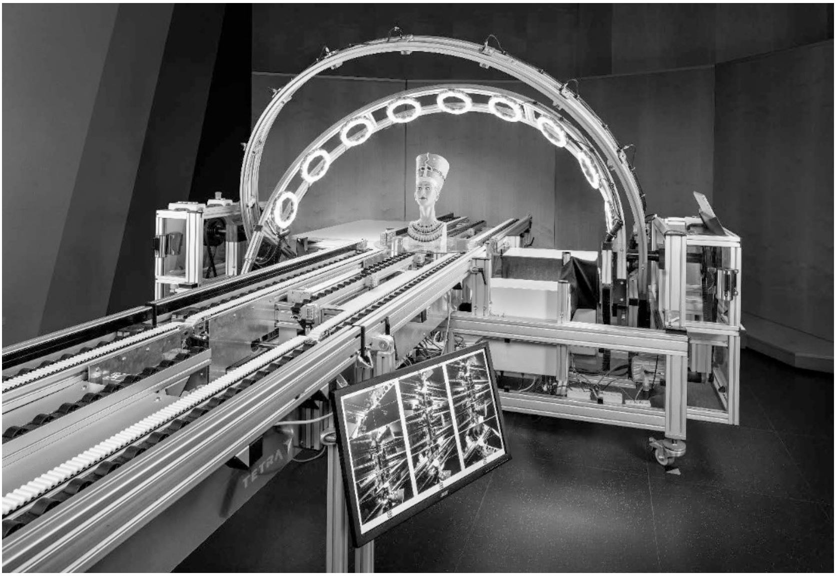


Fig. 3.2 CultArc3D: two hemispherical rotating arches, one with cameras, the other with ring light sources (Fraunhofer IGD)

As with the outer camera arch, the inner arch (hereinafter “light arch”) includes nine light sources mounted equidistantly. At the moment, all objects are captured in the visible light spectrum. However, for capturing more optically-complicated material, multispectral sensors or laser sensors can easily be integrated into the system, as can volumetric data capture sensors, x-ray tomography or MRI. One additional strength of the CultArc3D is the capture of optical material properties. To this end, both arches can be positioned anywhere in relation to one another, allowing every possible combination of light direction and photographic perspective for capturing an object’s upper hemisphere.

CultArm3D

Few fields present objects as varied in material and form or demands as high in terms of the precision and color-accuracy of digital replicas as the cultural heritage field. In order to guarantee the complete and accurate 3D reconstruction of any object it is thus important to position the scanner carefully with regard to the object’s surface geometry and the scanner’s measuring volume. The CultArm3D (see Fig. 3.3) is a robotic system for this kind of automated 3D reconstruction. It consists

Fig. 3.3 CultArm3D: lightweight robotic arm with a 24-megapixel camera mounted to the end effector next to a turntable with a white two-part folding studio box (Fraunhofer IGD)

of a lightweight robotic arm with a camera on its end effector and a turntable for the object to be scanned. It can operate either independently or in conjunction with the CultLab3D; in independent mode it is capable of capturing geometry and texture completely and independently.

The robotic arm selected has five degrees of freedom and is able to stably move loads of up to 2kg. It is collaborative and safe to operate in the presence of people, within a quasi-hemisphere of around 80cm. The robotic arm is equipped with a high-resolution (24 megapixel) camera and is synchronized with a diffuse lighting structure (static soft box setup). A two-part white studio box on the turntable ensures a consistent background, which is important for proper photogrammetric 3D reconstruction and the avoidance of incorrect feature correspondences.

Together with the CultLab3D, the CultArm3D serves as a second scanning station to the CultArc3D, which is the first scanning station on the conveyor. Unlike the CultArc3D with its large field of view on a fixed hemisphere around the object, the camera optics of the CultArm3D system are optimized for adaptive close-up detail images. These views, based on the 3D preview model of the first CultArc3D scan, are then planned so that additional features of the object will be captured. In this way, any remaining gaps or undercuts can be resolved and the quality and completeness of the scan can be improved locally. The number of additional planned views to meet specific quality criteria largely depend upon the complexity of the object's surface and texture. The automated planning of different numbers of views and the associated dynamic scanning time, depending on the complexity of the object, result in high throughputs of the CultLab3D digitization pipeline. When operating the CultArm3D in standalone mode separately from the CultLab3D assembly, the first round of scanning is carried out using generic camera angles that are only planned relative to the object's external cylindrical measurements such as height and diameter.

Given the limited workspace of the lightweight robotic arm, the task of capturing any kind of object up to 60cm in height and diameter is demanding. In some cases, camera views planned to provide optimal scan coverage and quality cannot be captured exactly, for reasons of safety or reachability. Nevertheless, in most cases it is possible to identify slightly modified, practical views (or series of views) that equally contribute to the quality of the scan.

3.1.2 Results, application scenarios, and future developments

The use of robot-operated 3D scanners for capturing uniform individual components has been standard in industry for a long time now. Cultural artifacts present a new challenge, however, due to their unique quality structures. Advances in 3D digitization and automation technology have now brought their cost-effective use within reach. For the first time ever, objects of different shapes and sizes can be digitized in large quantities and at high quality (see Fig. 3.4 and 3.5).

The digitization pipeline described above also offers a number of potentially useful applications in industry. Three-dimensional digitization of product portfolios for resellers, home improvement stores or mail order businesses is just one possible area of use. The long-term goal is to produce consolidated 3D models. These are digital copies of real objects which bring together the outputs from various measuring procedures in a single 3D model. A consolidated 3D model might, for example, amalgamate the data from a surface scan with those from a volumetric scanning procedure (e.g. CT, MRI, ultrasound) and a strength analysis. CultLab3D already has flexible modules for capturing 3D geometry, texture and material properties. But its underlying scanner design allows for actual as well as virtual enhancement

Fig. 3.4 Left: 3D reconstruction of a replica of Nefertiti – mesh; right: 3D reconstruction of a replica of Nefertiti – final color-calibrated 3D model (Fraunhofer IGD)

of the 3D models with data from the broadest possible range of scans including CT, ultrasound, confocal microscopy, terahertz imaging, and mass spectroscopy. This approach enables objects to be comprehensively investigated in their entirety, inside and out, and thus opens up new possibilities for monitoring, analysis and virtual presentation extending beyond the cultural sphere.

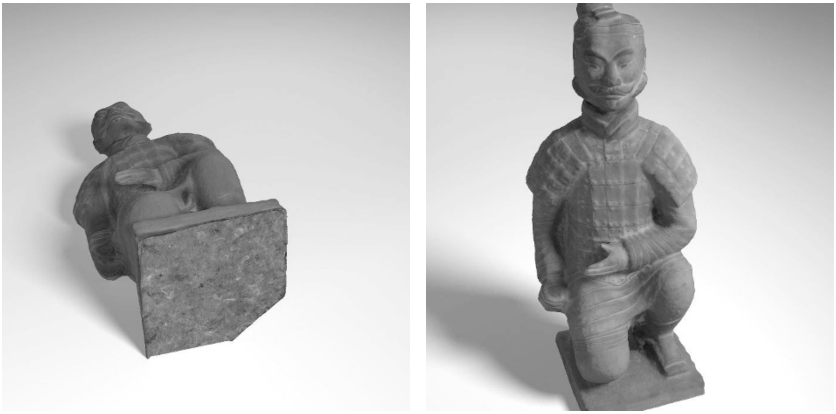


Fig. 3.5 3D reconstruction of a replica of a Chinese warrior; CultLab3D also captures the undersides of objects (Fraunhofer IGD)

3.2 Virtual and Augmented Reality systems optimize planning, construction and manufacturing

Today, Virtual and Augmented Reality technologies (VR/AR) are scaled for different computing capacities, operating systems and input and output options, from complete cloud infrastructures right through to head mounted displays. Along with scalability and platform diversity, mobile systems also bring completely new security requirements with them, since confidential data should be transmitted wirelessly, or CAD data should be visualized on mobile systems but not saved to them. Virtual and Augmented Reality technologies will only fulfill these requirements in future if they are based on web technologies, which are platform-independent and optimized for security. Against this backdrop, current research and development work in the VR/AR fields is closely linked to web technologies that provide enormous benefits, particularly for industrial uses.

3.2.1 Virtual Reality

Virtual Reality has been successfully utilized in European industry to make digital 3D data tangible for over 25 years. In the automotive industry and in aircraft and shipping construction in particular, digital mock-ups (DMUs) are replacing real and physical mockups (PMUs) in numerous areas of application.

Industrial VR applications use general manufacturing and DMU data for digital validation processes (e.g. assembly/development validation) in spite of the relatively high hardware and software costs of VR solutions.

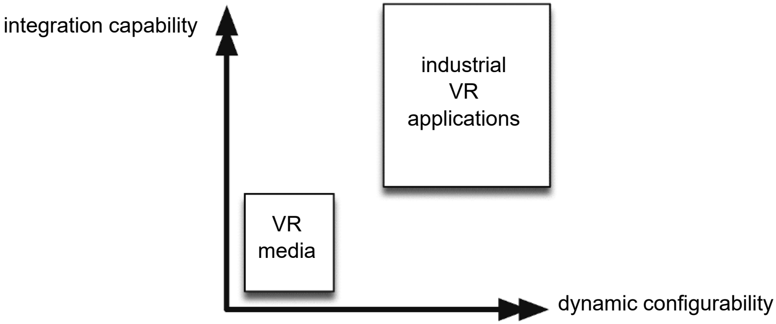


Fig. 3.6 Categorization of established VR media/gaming technologies and new industrial requirements (Fraunhofer IGD)

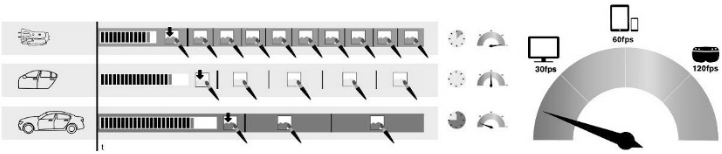


Fig. 3.7 "Input data sensitive" visualization issues with current standard VR processes; the image refresh rate increases with the size of the data, an unacceptable situation for VR/AR application. (Fraunhofer IGD)

Over the last five years, overall improvements in LCD and OLED technology have reduced the costs for VR headset to just a few hundred Euros. The associated VR games and media are available in the established online stores (e.g. Stream) and are undergoing a fierce pricing war. For industrial VR applications, no universal solutions have yet become established since the applications require a significantly higher capability of being integrated and dynamically configured. Classical solutions from the gaming market are not directly transferrable since the processes and data need to be adapted via manual configuration processes.

Dynamic configurations require the "smart" fully-automated adaptation of processes in order to be able to cope with the ever-widening distribution and explosion of data volumes. Current standard systems for industrial 3D data visualization use established document formats, and the processes used are "input data sensitive". As the volume of data grows, so does the visualization overhead.

Despite the constant increase in the processing power of modern graphics hardware, visualization of massive 3D data sets with interactive frame rates cannot be solved simply by increasing graphics memory and rasterization speeds. Problems such as visibility calculations for scene elements have to be accelerated by spatially hierarchical data structures. Instead of processing the highly complex geometry of individual elements in a scene, these data structures focus only on their bounding volume. Here the "divide and conquer" approach applies, based on recursive analysis of the tree structures produced.

Various concepts from different lines of research are available for structuring these kinds of hierarchies, for example collision detection or ray tracing. Whereas binary k-d trees, for example, split the allocated space in each node with a hyperplane, the structure of a bounding volume hierarchy is based on the recursive grouping of the bounding volumes of the elements in the scene. Hybrid processes such as bounding interval hierarchies connect these approaches in order to combine the advantages of each. These irregular approaches can be contrasted with structures such as regular grids or octrees where the space division strategy largely or completely

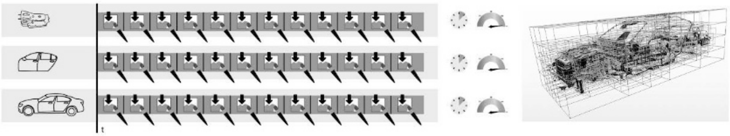


Fig. 3.8 Smart Spatial Services (S3) use a constant frame rate for globally budgeting visibility calculations and image generation. As with film streaming, a constant frame rate is the goal. (Fraunhofer IGD)

ignores the density of the space being divided. These different approaches are continuously being developed and optimized within the various fields of research.

The uniqueness of so-called “out-of-core technologies” lies in the fact that they do not save all their data in the main memory but load it on demand and on the fly from the secondary storage. In contrast to the automated virtual memory management of operating systems—which are normally not influenced by applications—programs here require complete control of memory management and data caching. Ideally, applications are supported in this by output-sensitive algorithms.

The focus of visualization here is less on 100% accuracy and/or completeness of the representation and more on guaranteeing a certain level of desired performance in terms of hard real-time requirements.

This can be achieved by the use of spatial and temporal coherence. Whereas spatial coherence is represented via the hierarchical data structures, temporal cohesion across frames has to be captured within the render algorithm. Hidden surface determination of the elements within the scene (“occlusion culling”) may for example be achieved via algorithms such as coherent hierarchical culling.

3.2.2 Augmented Reality

In the Industry 4.0 context, simulation and production processes are parallelized with the aim of guaranteeing optimal production quality via the comparison of target and actual outputs (cyber-physical equivalence). In the same way, the digital twin is used to feed the actual data back into an agile production planning process.

Augmented Reality (AR) processes are pertinent here for registering the target/actual differences in real time and visualizing them in superimposition to the captured environment. Augmented Reality processes have proven themselves in numerous areas of application in this context and are already finding applications in routine planning and monitoring processes:

- Augmented Reality shop-floor systems
  Today, Augmented Reality systems are already being offered as productive systems. Here, car mechanics are guided step-by-step through complex repair scenarios (see Fig. 3.9). The goal here is to tailor the Augmented Reality repair instructions exactly to the vehicle, considering the specific configuration and features.

- Augmented Reality-assisted maintenance
  With the same intention Augmented Reality systems are often combined with a remote expert component. Here, the camera images captured by the AR system are transmitted to an expert via the Internet, who can provide additional assistance in repair scenarios. This information may be combined with the AR system and the AR repair instructions, so that the remote expert component can simultaneously be used as an authoring system to supplement the repair instructions.

- Augmented Reality manuals
  Augmented Reality manuals for complex devices can provide an extensive graphical and hence language-independent tutorial for complex operating instructions. The manuals can be distributed via app stores for the various smartphone and tablet systems and thus easily updated and maintained (see Fig. 3.9).

Augmented Reality technology is in particular receiving significant attention because new hardware systems are being developed that facilitate entirely new interaction paradigms, completely revolutionizing working and procedural methods in production planning and quality control. Microsoft's HoloLens system is the pioneer here, integrating a multimodal sensor technology while also providing high image quality with an optical see-through system (see Fig. 3.10). However, this system also clearly demonstrates the limitations of the technology for professional use:

Fig. 3.9 Demonstrator of an Augmented Reality shop-floor system (Fraunhofer IGD)

- Content preparation overhead
  For current solutions developed for HoloLens, 3D models have to be reduced to a size suitable for AR with significant manual overhead. (The model size for the HoloLens system recommended by Microsoft is just 100,000 polygons; a standard automobile CAD model comprises approx. 80 million polygons.)

- Tracking with HoloLens
  In the SLAM-based (Simultaneous Localization and Mapping) pose estimation process utilized, tracking is initialized via gesture-based user interaction; that is, the models are oriented exactly as they are placed by the user via gesture. This gesture-driven initialization cannot be implemented by the user in a way that is sufficient for target/actual comparisons, for example.

- Tracking in static environments
  The tracking is based on 3D reconstructions of 3D feature maps and 3D meshes that are built during the application runtime. For this reason, the tracking has difficulties in dynamic environments such as when several individuals are using the HoloLens system at the same time, or when components to be tracked are moved.

Fig. 3.10 Demonstrator of Augmented Reality-assisted maintenance with HoloLens (Fraunhofer IGD)

- Tracking CAD models
  The SLAM-based tracking process is not able to differentiate between different objects to be tracked. It also cannot differentiate between what belongs to the object or to the scene background. The process is thus not suitable for target/actual comparison scenarios that are verifying an object's alignment in relation to a reference geometry.

- Gesture-based interaction
  Gesture-based interaction via the HoloLens is designed for Augmented Reality games and is not always suitable for industrial applications. For this reason, it should be possible to control the interaction with a tablet while the Augmented Reality visualization is executed on the HoloLens.

- Data storage on mobile systems
  Current solutions force (reduced) 3D models to be saved on the HoloLens. This can result in data security and data consistency processes being put at risk.

These restrictions in using HoloLens can only be compensated for if the algorithms are distributed across client server infrastructures. The key here is that the complete 3D models – which represent significant IP in industrial organizations – do not leave the PDM system and are saved exclusively on the server. Then the client only displays individual images transmitted via video streaming technologies. Alternatively, G-buffers are streamed, processed for rendering the current views. VR/AR technologies will only be able to meet these demands in future if they are built on web technologies and service architectures. The development of VR/AR technologies based on service architectures is now possible, as libraries are now available such as WebGL and WebCL, which facilitate powerful and plugin-free on-chip processing on the web browser. Web technologies as a basis for VR/AR applications in particular offer the following benefits for industrial applications:

- Security
  If VR/AR applications are run as web applications on the user's end device (smartphone, tablet, PC, thin client), then, in the best case scenario, native software components – which always entail potential insecurities – do not need to be installed at all.

- Platform dependence
  On the whole, web technologies can be used independent of platform and with any browser. In this way, platform-specific parallel development (for iOS, Android, Windows etc.) can be avoided as far as possible.

- Scalability and distribution
  By using web technologies, CPU-intensive processes can be well distributed over client server infrastructures. In doing so, the distributed application can be scaled not only to the computing power of the end device, but also to web connectivity, to the number of concurrent users, and to the volume of data required.

One example of industrial relevance in this context is the linking to the PDM system that centrally manages and versions all the relevant product data (e.g. CAD data, simulation data, assembly instructions). Central data storage is designed to ensure that the most current and suitable data version is always used for planning and development processes. In the field of VR/AR applications for target/actual comparisons in particular, correct versioning needs to be guaranteed. Using service architectures, VR/AR applications can be created that pull the current version from the PDM system when the application is started and, during data transfer, code it into geometric primitives that can be visualized in the web browser. To do this, 3D data is organized into linked 3D data schemas, which permit flexible division and use of the data in the service architecture.

3.2.3 Visualization using linked 3D data schemas

The large quantity of data and security requirements in automobile manufacturing exclude the full transfer of the entire CAD data to the client. For this reason, 3D visualization components must be able to adaptively reload and display relevant areas of the application. To do this, the 3D data is converted on the server side into a form optimized for the visualization. The key element here is the 3D data from CAD systems, made up of structural data (e.g. the position of a component in the room) and geometric data (e.g. a triangular mesh representing the surface of a component). Usually, these are arranged in a tree structure that can show the cross-references to other data and resources (see Fig. 3.11).

Fig. 3.11 Left: 3D data from CAD systems is usually depicted as a linked hierarchy of resources; right: from the point of view of the application, the 3D data is made up of a linked network of 3D data. (Fraunhofer IGD)

A typical PDM structural description may, for example, be mapped in the standardized STEP AP242/PLM XML formats, which reference geometries that can themselves be saved in the JT format. (These in turn reference additional JT subcomponents and may contain additional 3D data.) Development work is currently being carried out at Fraunhofer IGD on the instant3Dhub web-based service infrastructure for visualizing optimized data. Here, 3D data is stored in a linked 3D data network ("linked 3D data" or "L3D") that provides the complete structure and geometry transparently to all resource formats and is scalable and extensible via additional links (see Fig. 3.11 right).

Conversion between the resource description and the linked 3D data takes place via a transcoding process. Since the infrastructure was designed for fast, adaptive data reload it is equipped with corresponding delivery strategies. Here, the data is stored in a distributed 3D cache that is populated via transcoding jobs and transmitted via regular services.

The client-side application controls the display of the linked 3D data. instant3Dhub provides this access via a client-side API: webVis JavaScript API for integration into a browser application. The API communicates interactions and changes in camera position, initiating server-side visibility analysis and new links.

When the data is accessed, application-managed transcoding happens transparently; that is, the client only uses the linked 3D data, and the associated cache entry is only automatically generated on initial access. In order to speed up the access, the data can be completely transcoded at the point of publication, through successive access to the corresponding entries.

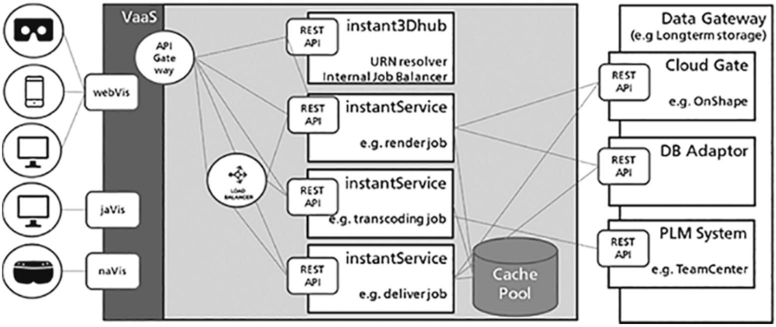


Fig. 3.12 Client/server architecture for 3D components: the client-side application uses the webVis API to access cache entries. The instant3Dhub service manages the transmission of cache entries and/or any transcoding required. A data gateway provides the connection to original 3D resources. (Fraunhofer IGD)

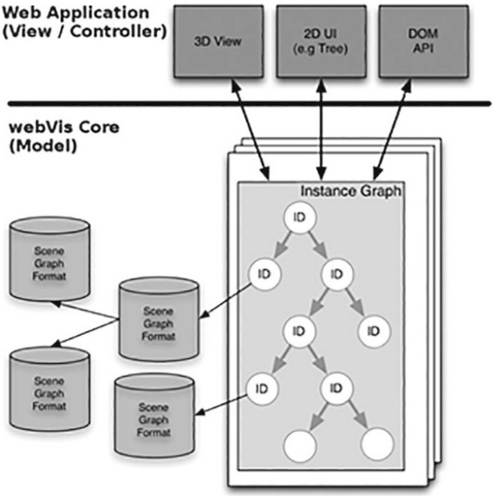


Fig. 3.13 Client-side view of the data model for 3D components. The web application loads and/or modifies entries of the instance graph and responds to user events. (Fraunhofer IGD)

Individual cache entries are identified via a naming scheme using Uniform Resource Name (URN) coding, so that they can be permanently identified on the application side regardless of location. The advantage of the 3D data network is that the storage location for data delivery can be changed without also changing the client application.

On the client side, the 3D components consist of a JavaScript application running on the web browser. The client-side JavaScript library, for managing the 3D components (webVis), offers a lightweight API for

- Adding or removing structural elements to and from the scene displayed,
- Reading and changing properties,
- Carrying out measuring functions, and
- Setting the visibility of (sub-)components.

When adding an element higher up the hierarchy, the entire hierarchy beneath is automatically displayed. Properties that may be changed include visibility (for showing or hiding elements) and color, for example. Alongside the direct 3D view, additional UI components (as web components) such as a tree view of the data structure or a bar for saving and loading current views ("snapshots") are also available.

The 3D component responds to user inputs (e.g. mouse clicks on a visible element of the 3D view) via events, i.e. a callback method is registered for the selected components. It is possible to register corresponding “listener” callbacks for a wide range of events and status changes of the graphical representation. For frequently needed client-side functionalities, a toolbox is available that allows different tools (e.g. clipping planes or screenshots) to be added to the application.

instant3Dhub offers a service-oriented approach to provide a unified 3D-as-a-service layer for application developers. For these applications, the infrastructure provides a bidirectional client interface. This interface is able to represent a range of different 3D data formats, including structural data and metadata, directly within an HTML element (e.g. <canvas> or <iframe>). To do this, a service-oriented service infrastructure was created that requires no explicit conversion or provision of replacement formats, but instead automatically creates the necessary containers for the various classes of end devices and keeps appropriate server-side cache infrastructures available.

3.2.4 Integration of CAD data into AR

Based on the foundation of instant3Dhub and WebGL-based rendering, special data containers were developed that allow 3D mesh data to be efficiently and progressively transferred over the network [8]. In contrast to classical out-of-core rendering processes [9], the processes presented here require no intensive pre-processing but can be used directly on the CAD data without any preparation [10]. As with NVIDIA’s proposed rendering-as-a-service process, this process is based on web infrastructures: it is designed to implement hybrid rendering processes that connect a powerful GPU cloud with client systems. However, where with NVIDIA the image data is highly efficiently compressed and transmitted [11], in instant3Dhub visibility-dependent triangles for rendering (the 3D mesh data for rendering) are used. To do this, alongside other services, new compression and streaming processes were developed [12] which initially only transmit the pre-decimal digits for each geometry to be streamed, in order to then successively send all the decimal places. Using this approach, a progressive rendering algorithm was implemented that initially visualizes the geometry with a limited degree of accuracy, which increases to complete accuracy with continued data transmission. These technologies were integrated into the instant3Dhub infrastructure – an infrastructure highly relevant to industrial customers which thus also enables connections to a PDM environment.

The availability of CAD models also allows model-based tracking, however. Here, the CAD data is used to generate reference models that are compared with the silhouettes recognized in the camera image. These model-based tracking procedures respond robustly to varying lighting conditions in unstable industrial lighting environments. Indeed, this also raises the question of how CAD models can be efficiently distributed to the output units for AR applications such as tablets, smartphones or HoloLenses, and how reference data can be generated for individual client applications.

3.2.5 Augmented Reality tracking

Augmented Reality processes are relevant for the development of intuitive-perceptive user interfaces and are being used to capture target/actual differences. Even so, the elemental core technology of Augmented Reality is the tracking technology, which allows the camera position relative to the viewed environment to be registered. Traditional approaches (e.g. marker-based or sensor-based approaches) are impertinent for industrial applications because the preparation overhead (measuring the markers, initializing via user interaction, etc.) is unprofitable or the processes involve significant drift. Feature-based SLAM processes, too, are irrelevant for industrial applications: they are dependent on the lighting situation and are not able to differentiate the tracked object from the background. Use of these processes is thus restricted to entirely static environments. The only approach suitable for industrial applications is thus model-based tracking, since it is able to create a reference between CAD data and the captured environment. Model-based processes do

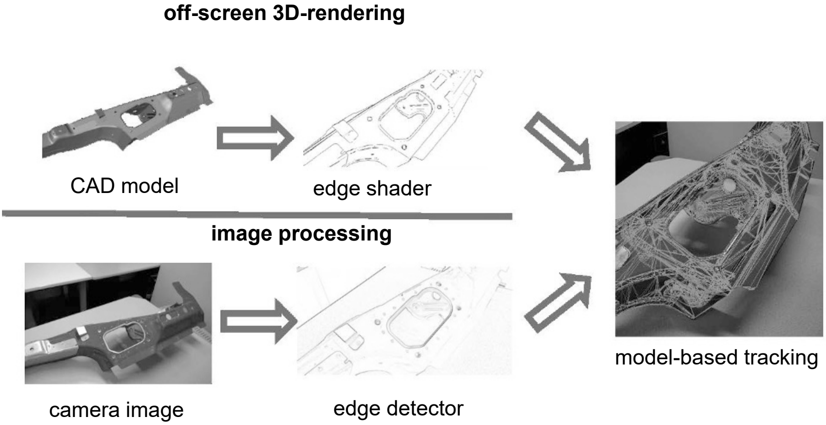


Fig. 3.14 In model-based tracking, edge shaders are used to render hypotheses from the CAD data that are then compared with the camera images and applied to the edge detectors. (Fraunhofer IGD)

not require user interaction for initialization and have no drift. The CAD geometry is permanently geared to the identified geometry. In addition, CAD nodes – which can be separated from one another in the structure tree, for example—can be tracked independently of one another. These are precisely the properties that are fundamental for Industry 4.0 scenarios in the area of target/actual comparisons and quality control.

Alongside the continuous tracking of the objects (frame-to-frame), the model-based tracking processes need to be initialized. During initialization, the real and virtual 3D objects are converted into a shared system of coordinates. To do this, initial camera positions are determined from which the initialization can be carried out. These positions for initialization are specified using the 3D models.

3.2.6 Tracking as a service

Just like the visualization functionalities, the tracking services are integrated into the instant3Dhub infrastructure. This enables new forms of efficient Augmented Reality usage to be implemented, via the same philosophy of data preparation and transmission. To this end, services are implemented for inferring and producing tracking reference models from CAD data that are transmitted via L3D data containers (see Fig. 3.15). Incorporating this into the instant3Dhub infrastructure will

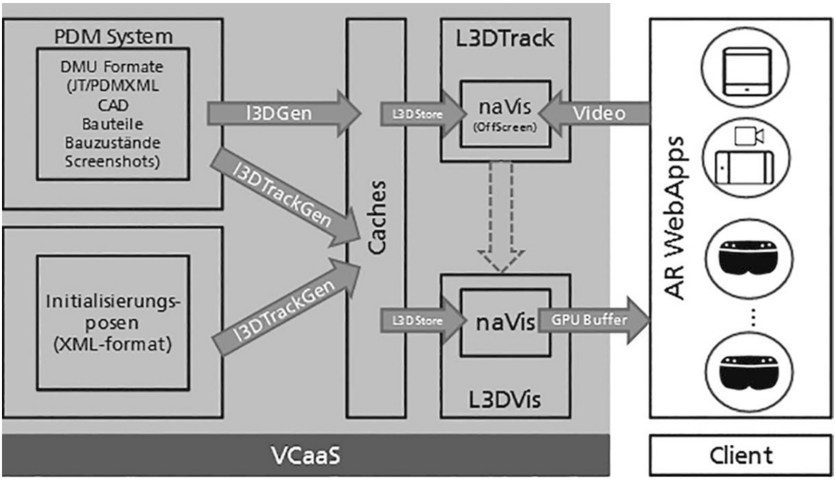


Fig. 3.15 Incorporation of tracking infrastructure into the VCaaS (Fraunhofer IGD)

enable new forms of load balancing for tracking tasks (distribution across client-side or server-side processes). However, the most important benefit of using the instant3Dhub infrastructure is that the reference models for the model-based tracking are generated directly from the CAD data at full resolution. This completely removes the need for expensive model preparation and reduction. To allow this, the instant3hub infrastructure provides efficient server-based rendering processes for large CAD models.

The model-based tracking processes in the instant3Dhub infrastructure use off-screen rendering processes to render the objects for tracking from the current camera position (tracking hypotheses). These off-screen renderings are intended to be used not only for tracking but also for identifying the state of construction. Which components have already been used in the current state, and which have yet to be added? These states of construction are to be displayed in screenshots and recoded each time for the tracking reference model.

With the help of the instant3Dhub infrastructure, the tracking service can be used in the following configurations according to parameters (network quality, client system performance, etc.):

- Hybrid tracking
  In this approach, image processing takes place on the client, while the tracking hypotheses are rendered on the server. The advantage of this approach is that the elaborate image processing is executed on the client, enabling minimal tracking latency. In this approach, however, a native tracking component must be installed on the client. The advantage of using the server as opposed to a purely native process is that the model data does not need to be saved on the client and the models for the model-based tracking do not need to be reduced.

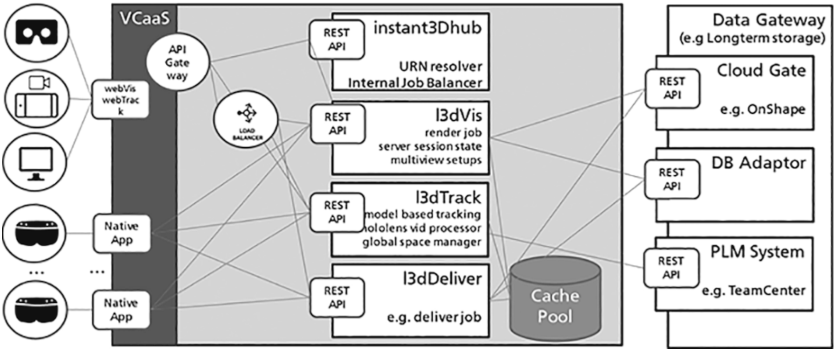


Fig. 3.16 Incorporation of tracking infrastructure into the VCaaS (Fraunhofer IGD)

- Server-side tracking
  For a server-side tracking process, the video data needs to be transmitted to the server infrastructure with minimal latency. The image processing takes place on the server and the camera positions calculated need to be transferred back to the client at a high frequency. This requires a video streaming component that must be implemented via a dedicated transmission channel (e.g. WebSockets). Server-based tracking requires a very good network connection and implementation in an asynchronous process so that positions can also be calculated asynchronously if there are bottlenecks in the communications infrastructure.

An instant3Dhub infrastructure with added model-based tracking thus takes shape as shown (see Fig. 3.16) with corresponding tracking services and the inference of reference models (I3dTrack). All the processes for real-time operation of an AR shop floor system are thus combined in this infrastructure.

Sources and literature

[1] European Commission: Commission recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation (2011/711/EU) [online], available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:283:0039:0045:EN:PDF (accessed 5 June 2016).
[2] European Commission: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A Digital Agenda for Europe (COM(2010)245 final) [online], available from: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52010DC0245&from=en (accessed 5 July 2016).
[3] Article 3.3 of the Treaty of Lisbon, Treaty of Lisbon, Amending the Treaty on European Union and the Treaty Establishing the European Community, 17.12.2007 (2007/C 306/01), in: Official Journal of the European Union, C 306/1 [online], available from: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:C:2007:306:TOC (accessed 10 September 2015).
[4] Keene, Suzanne (Ed.) (2008): Collections for People. Museums' stored Collections as a Public Resource, London 2008 [online], available from: discovery.ucl.ac.uk/13886/1/13886.pdf (accessed 29 June 2016).
[5] Santos, P.; Ritz, M.; Tausch, R.; Schmedt, H.; Monroy, R.; De Stefano, A.; Fellner, D.: CultLab3D – On the Verge of 3D Mass Digitization, in: Klein, R. (Ed.) et al.: GCH 2014, Eurographics Workshop on Graphics and Cultural Heritage. Goslar: Eurographics Association, 2014, pp. 65-73.
[6] Kohler J., Noell T., Reis G., Stricker D. (2013): A full-spherical device for simultaneous geometry and reflectance acquisition, in: Applications of Computer Vision (WACV), 2013 IEEE Workshop (Jan 2013), pp. 355-362.

[7] Noell T., Koehler J., Reis R., and Stricker D. (2015): Fully Automatic, Omnidirectional Acquisition of Geometry and Appearance in the Context of Cultural Heritage Preservation. Journal on Computing and Cultural Heritage – Special Issue on Best Papers from Digital Heritage 2013 JOCCH Homepage archive, 8 (1), Article No. 2.

[8] Behr J., Jung Y., Franke T., Sturm T. (2012): Using images and explicit binary container for efficient and incremental delivery of declarative 3D scenes on the web, in ACM SIGGRAPH: Proceedings Web3D 2012: 17th International Conference on 3D Web Technology, pp. 17–25

[9] Brüderlin B., Heyer M., Pfützner S. (2007): Interviews3d: A platform for interactive handling of massive data sets, in IEEE Comput. Graph. Appl. 27, 6 (Nov.), 48–59.

[10] Behr J., Mouton C., Parfouru S., Champeau J., Jeulin C., Thöner M., Stein C., Schmitt M., Limper Max, Sousa M., Franke T., Voss G. (2015): webVis/instant3DHub – Visual Computing as a Service Infrastructure to Deliver Adaptive, Secure and Scalable User Centric Data Visualisation, in: ACM SIGGRAPH: Proceedings Web3D 2015 : 20th International Conference on 3D Web Technology, pp. 39-47

[11] Nvidia grid: Stream applications and games on demand. http://www.nvidia.com/object/nvidia-grid.html

[12] Limper M., Thöner M., Behr J., Fellner D. (2014): SRC – A Streamable Format for Generalized Web-based 3D Data Transmission, in: ACM SIGGRAPH: Proceedings Web3D 2014 19th International Conference on 3D Web Technology, pp. 35-43

Video Data Processing

4

Best pictures on all channels

Dr. Karsten Müller · Prof. Heiko Schwarz · Prof. Peter Eisert ·
Prof. Thomas Wiegand
Fraunhofer Institute for Telecommunications,
Heinrich-Hertz-Institute, HHI

Summary

More than three quarters of all bits transmitted today over the consumer Internet are video data. Accordingly, video data is of major importance for the digital transformation. The related field of digital video processing has played a key role in establishing successful products and services in a wide range of sectors including communications, health, industry, autonomous vehicles and security technology. Particularly in the entertainment sector, video data has shaped the mass market via services like HD and UHD TV or streaming services. For these applications, efficient transmission only became feasible through video coding with methods of efficient compression. In addition, production systems and processing techniques for highly realistic dynamic 3D video scenes have been developed. In these key areas, the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute, HHI is playing a worldwide leading role, in particular in the key areas of video coding and 3D video processing, also through successfully contributing to video coding standardization.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_4

4.1 Introduction: The major role of video in the digital world

Video data has become a central topic in the digital world. This was achieved through major contributions in the related research field in video processing during the last decades, e.g. by establishing a number of successful systems, business sectors, products and services [35], including:

- Communications: Video telephony, video conferencing systems, multimedia messenger services such as video chats are now used in both business and home environments.
- Health services: Digital imaging techniques allow displaying and processing high-resolution computed tomography scans. Here, virtual models for preparing and supporting surgery are reconstructed from medical data, and automatic image analysis methods of such data are carried out to assist in diagnosis.
- Industry: For automated process monitoring, quality and production control, cameras are used as sensors. Here, video analysis methods derived from pattern recognition are used to quickly and automatically verify whether products in a production line exactly meet the specifications. Digital video data has become even more important since the introduction of the Industry 4.0 postulate, leading to increased production automation and modeling of production processes in the digital world.
- Vehicles and logistics: Video data from camera sensors is being utilized for automatic traffic routing. In logistics, this method has been used for a long time to implement fully automated cargo roads with computer-guided vehicles. For vehicles, self-driving systems are being developed for some years, which analyze data from all sensors for automated vehicle guidance All areas of security technology use optical surveillance systems and thus video data also plays an important role here.
- And finally, the entertainment industry as a mass market is being shaped by the primary role of video data. Here, entire business areas have been successfully developed, including television broadcasting in high-definition (HD) and ultra-high-definition (UHD), with resolutions of 3840 x 2160 pixels and above; mobile video services; personal recording and storage devices such as camcorders; optical storage media such as DVDs and Blu-ray discs; video streaming services as well as Internet video portals [57].

For achieving global distribution, international standards play a major role for the definition of video formats, image acquisition, efficient video coding, transmission, storage, and video data representation. This global role of digital video data is underlined by a study from Cisco [5], which shows that 73% of all bits transmitted over the Internet in 2016 were video bits, and it is further expected that this increases to 82% by 2021.

Digital video data represents projections of the visual world in digital form. The creation of this data requires image capturing or recording, followed by digitization. In real-world-capturing, similar principles as in the human eye are applied: Light enters the camera through the aperture, is bundled by a convex lens and hits a light-sensitive medium – a film or digital photo sensor. By this procedure, the real physical world is perceived in its familiar form, as the camera aperture only allows photons from the direction of the objective with small angular deviations. The subsequent light bundling produces a visible representation of the real world, similar to what we see with our eyes. With some simplifications, the brightness level of the resulting image is determined by the respective number of entering photons, and the color from the respective wavelengths. In order to advance from individual pictures to moving pictures or videos, pictures must be taken at particular time intervals. Due to the visual persistence of the human eye, a series of individual images displayed at a high enough rate (more than 50 images per second) is perceived as a moving picture/video.

In order to transform a recorded film to digital video data, the individual images are first discretized, i.e. spatially sampled. For analog recordings, this is done explicitly; for recordings using a digital camera sensor, the maximum number of light sensors in the sensor array already provides implicit discretization, e.g. an HD sensor records an image resolution of 1920 x 1080 pixels. Next, the color and brightness levels of the individual pixels are discretized in order to be represented in digital form. Typically, the three color values red, green and blue are discretized or quantized into 256 different values for each pixel. Each color value is then represented in binary format by 8 bits ($256 = 2^8$) and finally, the entire digital video can be easily processed, saved and coded by computer systems.

As a result, video data can be specified by formats that describe the properties and settings of the recording and digitization stages. E.g., the format "1920 x 1080 @ 50fps, RGB 4:4:4, 8bit", specifies a digital video with a horizontal and vertical resolution of 1920 and 1080 pixels respectively, a refresh rate of 50 images per second, in RGB color space, with the same resolution of all color components (4:4:4) and each quantized into 8 bits.

Video data has achieved global significance and worldwide distribution through digitization and the rise of the Internet. Digitization led to video data of new previously unseen quality, since video signals could now be processed efficiently and their data volumes drastically reduced via coding methods. Across the various distribution channels such as television and Internet broadcasting, via wired as well as wireless mobile channels, video data has become the most transmitted and consumed type of data. The overall volume of video data transmitted is growing faster than the capacity of transmission networks, meaning that – within the video processing chain from production to display and consumer playback – compression plays a predominant role. For the latter, international video coding standards are developed by the ITU VCEG (International Telecommunications Union – Visual Coding Experts Group, officially ITU-T SG16 Q.6) and ISO/IEC-MPEG (International Organization for Standardization/International Electrotechnical Commission – Moving Picture Experts Group – officially ISO/IEC JTC 1/SC 29/WG 11) in particular. Both standardization bodies often cooperate in joint teams and integrate improved compression methods into each new generation of standards. This enables video data to be transmitted at the same levels of quality with significantly lower data rates [32]. Notably, some applications have only become possible due to efficient video compression. E.g., the H.264/MPEG-4 AVC (Advanced Video Coding) standard [1] [23], enabled the wide distribution of high-resolution television (HDTV) and video streaming services via DSL (Digital Subscriber Line). Furthermore, DVB-T2 could only be commenced in Germany due to the successor standard H.265/MPEG-H HEVC (High-Efficiency Video Coding) [17] [47], as it enabled transmitting HD television with the same visual quality at the available (lower) data rate.

The digitization of video data has also initiated new areas of application, which extend two-dimensional video data into the third dimension in various ways. Here, fields of research have developed that target 3D object and scene modeling, using natural data captured with several cameras. In the field of computer graphics, powerful computers also enabled purely synthetic scene modeling, where entire animated films are modeled on the computer. Collaboration between the two fields has created mixed reality scenes that contain computer-animated graphics such as wireframes as well as natural video data. Key aspects here are accurate 3D perception of scenes and easy navigation. Accordingly, VR headsets were developed, which depict scenes in high-quality 3D – here, pairs of stereoscopic images, separately for the left and right eye – and which allow viewers to navigate within the scene much more naturally by moving or turning their head. In recent years, the field of virtual reality has further developed towards augmented reality (AR) and is likely to spawn application in various fields. These include, for example, medical technology, where computed tomography video data is combined with synthetic 3D models in order to improve diagnoses and surgery techniques. In architecture, building plans are increasingly produced virtually and can thus also be inspected virtually, allowing for more efficient planning of the real building process. Accordingly, a global market is currently also developing in this field, and all these areas add to the global importance of digital video data.

4.2 Video processing at Fraunhofer Heinrich-Hertz-Institute

The growing significance of digital video data has resulted in new technological challenges and market opportunities. In response, a video processing research section was formed at the Heinrich-Hertz-Institute (HHI) in 1989 within the Image Processing Department, in order to carry out early basic research, technological development and video standardization. In the last 15 years in particular, the video processing research section, now within the Heinrich-Hertz-Institute as an institute of the Fraunhofer Gesellschaft, has been playing a leading role in the field of video coding worldwide. In the academic field, this is expressed in a variety of renowned publications, guest lectures and editorial responsibilities for international journals and conferences. In addition, the institute has published an extensive two-part monograph on source coding [56] and video coding [57]. Important technologies for video coding have been developed and integrated into various standards. These include H.264/MPEG-4 AVC [59] along with its extensions SVC (Scalable Video Coding) and MVC (Multi-View Video Coding), successor standards H.265/MPEG-H HEVC [47] and their extensions for scalability, multi-view and 3D (SVHC [4], MV-HEVC [50] and 3D-HEVC [28][50]). Besides the technological co-development of standards, Fraunhofer HHI has been continuously involved in the management of relevant standardization committees. The acquired expertise is exploited commercially in a number of public and private projects. In parallel to the development of video coding technology, Fraunhofer HHI has also developed efficient transmission methods [16][39][40].

Based on its longstanding experience, Fraunhofer HHI has acquired an equally prominent position in the field of computer graphics and computer vision. Here, new methods for movie and television technology have been developed, allowing three-dimensional video content to be created, integrated and displayed in mixed reality scenes. Fraunhofer HHI has also carried out pioneering work in the field of synthetic and natural 3D video and computer graphics compression, as well as standardization work for dynamic wireframe models for MPEG-4 AFX (Animation Framework Extension).

In the video processing section, systems for 3D production such as the Stereoscopic Analyzer (STAN) for supporting depth control in 3D video content production have been created. Furthermore, technologies for video panorama recording with video camera systems, and automatic real-time stitching systems were developed in order to avoid visible transitions between the individual recording cameras in the panorama. By combining methods from the areas of computer vision, computer graphics and visual computing, new solutions for a broad range of applications in the fields of multimedia, augmented reality, medical imaging, and security could be developed. One example is the processing of video-based animated 3D models that are captured and reconstructed from real people and led to virtual 3D interactive films. With this, Fraunhofer HHI is also taking a leading role in international research and development in the new field of immersive 360° and VR technologies.

4.3 Compression methods for video data

Section 4.1 described that video services such as UHD TV and Internet streaming are only possible by coding standards with efficient video compression methods such as H.265/MPEG-H HEVC. As an example, we first consider the required rate for an uncompressed digital video signal with a high-enough temporal resolution of 50 images per second, each sent at UHD resolution of 3840 x 2160 pixels. This equals around 414 million pixels per second (50 images/s x 3840 x 2160 pixels/image). Each pixel represents a color value composed of the three components red, green and blue. At a moderate color resolution/quantization of 8 bits per color value, 24 bits are required for each pixel. This results in a bit rate of 10 Gbit/s for an uncompressed UHD video signal. In contrast, the available data rate for a UHD video is typically 20 Mbit/s. Uncompressed UHD signals are thus 500 times larger than the available transmission channel permits at maximum.

The above example thus requires a video compression method that compresses the original video to a size of 1/500 while maintaining a high video quality with no visible distortion. This leads to the formulation of a key requirement: An effective video compression method must provide the highest possible video quality at the lowest possible bit rate. This requirement can be found in every generation of video coding standards and is achieved via the general rate distortion optimization (RD optimization) in Eq. 4.1 [48] [58]:

$$J_{\min} = \min(D + \lambda R). \qquad \text{Eq. 4.1}$$

Here, the required rate R is added to the distortion D, where the additional parameter λ weights between the two variables [8]. Here, D is the deviation of a reconstructed video segment from the original segment and is inversely proportional to video quality. That is, the smaller D, the higher the video quality. Thus, the optimization requirement means achieving the lowest possible value of R with the lowest possible distortion D. This is achieved by minimizing the weighted sum in Eq. 4.1, and thus by finding the optimal Euler-Lagrange functional $J_{\min}$. This general formulation is used as a specific optimization for the choice of optimal coding mode, as described below. According to the area of application, a maximal video/transmission rate or maximal distortion/minimal quality may be specified. In the first case, an optimal video compression method would aim for the best possible quality (i.e. lowest distortion) at a given rate, in the second case, a minimal rate results from a given level of quality.

In order to achieve the required compression ratio for video signals, fundamental statistical analyses of image signals were first conducted [57], in particular including research on human perception [33]. The found properties were analyzed and applied to video coding. First, pixels are no longer displayed in RGB color space but in a transformed color space, better adapted to the human perception. Here, the YCbCr color space is used in particular, where the Y components contain the luminance information and the Cb and Cr components contain the chrominance information as a color difference signal between B and Y, and R and Y respectively. With respect to the different sensitivity of the human eye for luminance and chrominance, a data reduction can be achieved by subsampling the Cb and Cr components. In this way, a YCbCr 4:2:0 color format can be used for video coding, where the Y component of a UHD signal still contains 3840 x 2160 luminance pixels, while the chrominance resolution is reduced to 1920 x 1080 pixels for both Cb and Cr. For image representation and backward transformation into the RGB color space, one chrominance pixel from the Cb and one from the Cr signals are assigned to a quadruple of 2 x 2 luminance pixels from the Y signal. Subsampling the Cb and Cr signals already produces a reduction of the uncompressed data rate by a factor of 2, since each Cb and Cr only contain a quarter of the number of luminance pixels. In the case of video coding, it has also been shown that all texture details and fine structures are located in the luminance signal, while the chrominance signals have much less detail and can thus be higher compressed.

Additional statistical methods concentrated on natural video, i.e. camera recorded video data. Such videos can be rather different, as they show a great variety of various scenes and thus also feature different colors, color distributions as well as different motions and motion distributions. Thus, a filmed scene of a fleeing group of animals, with additional camera panning and zooming, shows a completely different pattern of color and motion in comparison to an anchor person in front of a static camera. Despite the differences in content of natural video sequences, natural video have some key similarities: within the different objects in a scene, neighboring pixels share a high local color similarity. This similarity exists in the spatial neighborhood of each frame, as well as in the temporal neighborhood between successive images. In the latter case, the local motion of an object must be considered in order to identify image areas between temporally neighboring images with high similarity [37].

Due to the similarity between neighboring pixels, lower data rates can already be achieved by taking the difference between neighboring pixels or blocks of pixels.

As a result, difference values between spatially or temporally neighboring pixels are further processed, instead of original color values of each pixel. Since the difference values are much smaller due to the high level of similarity in large areas of a video, less than the initially discussed 8 bits per color value are required. This data reduction is implemented in current video coding methods by taking the difference between an original image area/image block $\boldsymbol{s}[x, y, t]$ at spatial position (x, y) and temporal position t, and a corresponding estimated area $\hat{\boldsymbol{s}}[x, y, t]$ (see Fig. 4.1). Digital image signals are considered as PCM signals (pulse code modulated signals), such that the usage of difference signals leads to a differential PCM signal. Accordingly, the coding structure with a DPCM loop forms one of the key technology of modern hybrid video encoders (structure illustrated in Fig. 4.1).

The second essential key technology in compression is transform coding (see Fig. 4.1) where an image block is split into its harmonic components or 2D basis functions of different frequencies [1]. The original image block is then represented by the weighted sum of its 2D basis functions. The weights are called transform coefficients and their number is identical to the number of original pixels in a block, given the specific image transforms in video coding. The reason for the transformation efficiency in video coding can be found again in the natural images statistics, which reveals a high statistical dependency between neighboring pixels. After trans-

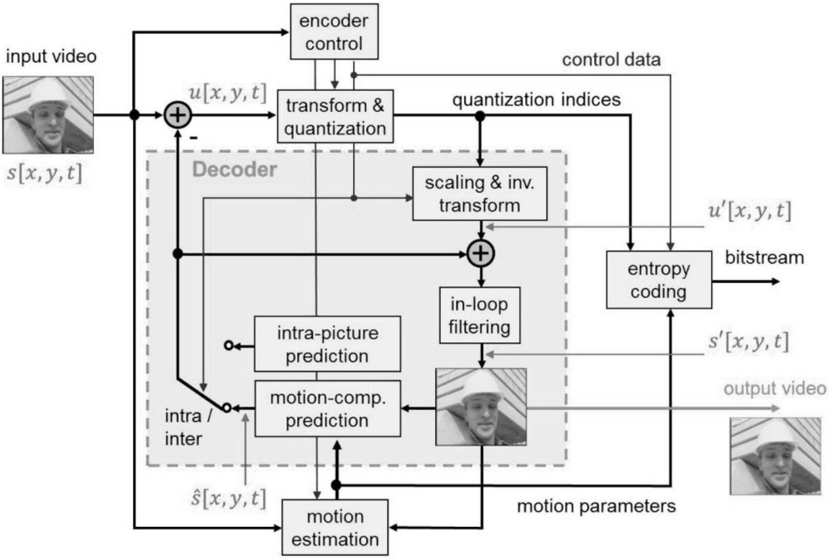


Fig. 4.1 Block diagram of a hybrid video encoder (Fraunhofer HHI)

formation, an image block can be represented by a small number of transform coefficients, which concentrate the signal energy. In extreme cases, an entire image block with homogenous color values can be represented by a single coefficient, which also equals the mean value of the entire block. Through subsequent quantization, only the most important coefficients remain, i.e. the once with the largest absolute values. This ensures that the maximum possible signal energy for a given bit rate is retained in the coded data stream and that the highest possible video quality at a given data rate is achieved after decoding.

Finally, the data is further reduced via entropy coding of the quantized transform coefficients. Here, a variable length code is employed that uses short code words for very frequently occurring values or symbols, and long code words for rarely occurring symbols. This way, further bit rate savings can be achieved, depending on the frequency distribution of symbols.

The basic structure for a classic hybrid video encoder with DPCM loop, transform coding and subsequent entropy coding, as shown in Fig. 4.1 has been described so far. To further exploit image and video similarities, additional video coding methods are used, as described in the following. Firstly, a video sequence is processed image-by-image in coding order. This may differ from the actual temporal image order, however allows using additional temporally succeeding images for good signal prediction, utilizing forward and backward prediction. Each frame is divided into image blocks $\boldsymbol{s}[x, y, t]$ that enter the coding loop. For each image block $\boldsymbol{s}[x, y, t]$, a prediction block $\hat{\boldsymbol{s}}[x, y, t]$ is calculated. For intra-predicted images (I-pictures), $\hat{\boldsymbol{s}}[x, y, t]$ is predicted exclusively from neighboring blocks of the same image. For inter-predicted images (P- and B-pictures), $\hat{\boldsymbol{s}}[x, y, t]$ can also be calculated from temporally preceding or succeeding images by means of motion-compensated prediction. In order to make optimal use of temporal similarity, motion estimation is carried out between the currently coded and a reference block, which provides the best prediction. As a result, a 2D motion vector with a horizontal and vertical component is obtained, that describes the estimated motion of the block between the current and reference image. Motion compensation is then carried out using the motion vector for good temporal prediction in $\hat{\boldsymbol{s}}[x, y, t]$. In each case, the predicted block $\hat{\boldsymbol{s}}[x, y, t]$ is subtracted from the original block $\boldsymbol{s}[x, y, t]$ in the coding loop, and the difference/residual signal $\boldsymbol{u}[x, y, t]$ is calculated. To identify the optimal predictor $\hat{\boldsymbol{s}}[x, y, t]$, the special rate distortion optimization in Eq. 4.2 [55] is used.

$$\mathbf{p}_{opt} = \arg \min_{\mathrm{p}} (D(\mathbf{p}) + \lambda \cdot R(\mathbf{p})). \qquad \text{Eq. 4.2}$$

For this, a range of coding modes $\mathbf{p}$ of the differently predicted $\hat{\boldsymbol{s}}[x, y, t]$, corresponding to distortion $D(\mathbf{p})$ and bit rate $R(\mathbf{p})$ are tested in order to find the optimal coding mode $\mathbf{p}_{\mathrm{opt}}$ to minimize the rate distortion functional [55]. These include different intra coding modes from the spatially neighboring blocks as well as different inter coding modes of temporally neighboring and motion-compensated blocks. To identify the optimal mode, $D(\mathbf{p})$ is calculated as the mean squared error between $\boldsymbol{s}[x, y, t]$ and $\hat{\boldsymbol{s}}[x, y, t]$, and thus as the variance of $\boldsymbol{u}[x, y, t]$. For the corresponding rate $R(\mathbf{p})$, the number of bits required for coding the block with the corresponding mode $\mathbf{p}$ is calculated (both for coding the transformed and quantized residual error signal and for the motion and signaling information [36]). Finally, the rate is weighted with the Lagrange parameter λ, which depends on the quantization selected [54], and finally guarantees the optimal coding mode $\mathbf{p}_{opt}$ for different bit rates.

In the next step within the coding loop, transform coding of the residual signal $\boldsymbol{u}[x, y, t]$ is conducted, e.g. via integer versions of the discrete cosine transform (DCT) [2]. The subsequent quantization and scaling of the transform coefficients is controlled via a quantization parameter (QP) that controls the rate point of the video encoder. Finally, the resulting quantization indices, motion vectors and other control information are coded losslessly using entropy coding. For this, CABAC (context-adaptive binary arithmetic coding) has been developed [22][49] as a powerful tool of entropy/arithmetic coding and integrated into both families of standards (AVC and HEVC).

As shown in the lower section of the coding loop in Fig. 4.1, the coded signal is reconstructed block-by-block as $\boldsymbol{s}'[x, y, t]$, to obtain a new prediction signal for the next loop cycle. To do this, the transformed residual signal is inversely scaled and transformed in order to obtain the reconstructed version of the residual signal $\boldsymbol{u}'[x, y, t]$. Subsequently, the current prediction signal $\hat{\boldsymbol{s}}[x, y, t]$ is added. For improved picture quality, a filter is applied to avoid visible block boundaries in the reconstructed image (in-loop filtering). Finally after this filtering, the reconstructed image is created. The encoder also contains the decoder (shaded gray in Fig. 4.1), thus knows the quality of the reconstructed video and can use it to optimize the compression method.

This describes the basic principles and working methods of modern video coding methods, in particular those of the AVC and HEVC standards. For details and precise descriptions of all the tools, please refer to the overview and description literature for AVC [23][45][46][52][59] and HEVC [12][15][20][21] [30][31][38][43] [47][49][51].

Although the same basic techniques have been used in all video coding standards since H.261 [60], the standards differ in key details, which led to a continuing increase in the achievable coding efficiency from one standard generation to the next. The majority of improvements here can be accounted to an increase in the supporting methods for coding an image or block. These include, among others, the number of supported transform sizes; the number of partitioning options and block sizes for intra prediction and motion-compensated prediction; the number of supported intra prediction modes; the number of usable reference images; the accuracy of the motion vectors coded, etc. Additional coding efficiency gains were achieved via an improvement in entropy coding and introduction of different in-loop filters.

As an illustration of the development of video coding, Fig. 4.2 shows a comparison of the coding efficiency of the most recent video coding standard H.265/ MPEG-H HEVC [17] versus its preceding standards H.264/MPEG-4 AVC [1], MPEG-4 Visual [6], H.263 [61] and MPEG-2 Video [14], for two test videos. For fair comparison, the same encoding control concept was used for all standards based on the Lagrange technique described in Eq. 4.2 above. For the first test video, *Kimono*, all encoders were configured such that an entry point at every second was available in the bitstream, from which decoding could start as required for streaming and broadcasting applications. The second video, *Kristen & Sara*, represents an example for videoconferencing applications; all of the images were coded in original recording order to provide as little delay as possible between

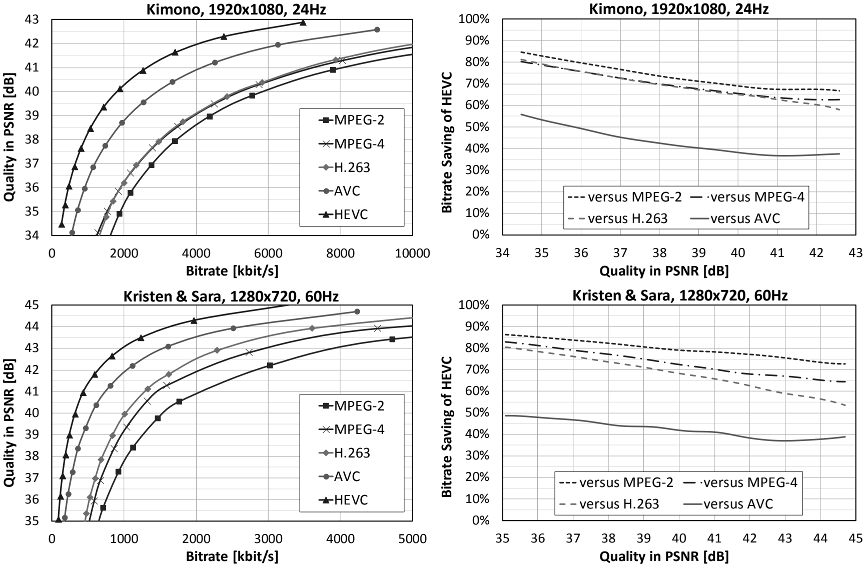


Fig. 4.2 Coding efficiency of international video coding standards for two selected test sequences. *Left:* reconstruction quality as a function of the bit rate; *Right:* bit rate savings of current HEVC standard in comparison to the different predecessor standards, source [32] (Fraunhofer HHI)

sender and receiver. For the results shown, PSNR (peak signal-to-noise ratio) was used as quality measure, defined as the mean squared error (mse) between original and reconstructed images, and calculated as PSNR = 10 $\log_{10}$ (255^2/mse). While the curves in Fig. 4.2 *left* compare the quality/bit rate functions, the graphs in Fig. 4.2 *right* show bit rate savings achieved with HEVC versus the predecessor standards for a given video quality (PSNR). In order to process a video at a particular video quality, the HEVC standard adopted in 2014 only requires 20–30% of the bit rate, necessary for the MPEG-2 standard from 1995. The bit rate savings for the same subjective quality, i.e. the quality as perceived by a viewer, are generally even larger [32].

In parallel to the highly successful standards for 2D video coding, extensions to AVC and HEVC were also specified, facilitating scalability [4][39][40][42], larger bit depths [10], larger color dynamic ranges [11], efficient coding of multiple camera views [26][50][52], and additional use of depth data [25][28][50][52].

4.4 Three-dimensional video objects

With the global distribution of digital video data, not only new markets for classical two-dimensional video data developed. Moreover, new research areas and technologies for producing and displaying 3D video scenes [7][24][27][44] have emerged. In contrast to classical video, where one dimension of the actual scene disappears due to projection onto the 2D image plane of the camera, the entire geometry of the environment is captured and represented via suitable 3D models. Thus, the choice of a viewing point for the scene is no longer determined by the recording method but can actually be freely selected by the user afterwards. By rendering the 3D scene to a virtual camera position, arbitrary camera and viewing trajectories become possible, as well as interactive navigation within the scene, where the viewer can freely select regions of interest. Additionally, by producing separate images for the viewer's left and right eye, a 3D impression of the observed scene can be created and thus the perception of the real spatial scene structure is improved.

The demand for producing three-dimensional video scenes has been driven recently both by the development of new virtual reality (VR) headsets that achieve improved immersion and natural viewer navigation in the virtual scene as well as by higher-quality 3D images. Users can now navigate scenes far more naturally by means of their own motion and head movements and better merge with the scene. This facilitates new multimedia applications such as immersive computer games, virtual environments (such as museums and tourist sights), and even innovative movie formats such as interactive films. Instead of moving in a purely virtual world, the connection of virtual 3D objects with the real scene in augmented and mixed reality applications is also possible, driven by technological advances in AR headsets (e.g. Microsoft HoloLens) or see-through applications using smartphones and tablets. Here, new content such as supplementary information and virtual 3D objects are inserted into the viewer's field of vision and registered in the real scene with the correct perspective. This technology will create new assistance systems and user support for sectors like health (endoscopy and microscopy), industry (production and maintenance), and mobility (driver assistance).

Currently, the high-resolution capture of three-dimensional dynamic environments is a challenge. The development of 3D sensors is further progressing, however the sensors are often less suitable for dynamic objects (e.g. scanners) or offer limited spatial resolution (e.g. time of flight sensors). Passive photogrammetric approaches, on the other hand, have gained importance due to increased camera resolutions, falling prices, ubiquitous availability and are capable of delivering high-quality results [3]. In multi-camera 3D systems, the scene is captured using two (stereo) or more (multi-view) cameras simultaneously from different viewing angles. Then, for each pixel in a reference image, corresponding pixels in the other images are identified, which originate from identical surface position of the real object. The displacement of the corresponding pixels between the two images is called "disparity" and can be converted directly to depth information from the camera plane to the real object point, since both variables are inversely proportional. The larger the displacement between two pixels, the closer the object point to the camera. As a result, a depth value can be assigned to each camera pixel. From this, a 3D point cloud is produced that represents the part of the object surface that is visible from the direction of the camera. By fusing the surfaces from different camera pairs, the entire scene can be reconstructed.

Fraunhofer HHI has developed numerous methods for analyzing and estimating 3D video objects in the sectors of multimedia/film [13][44], telepresence [53][18], health [41] and industry [35]. In the following, we describe 3D person reconstruction to illustrate the production of dynamic 3D video objects using multi-camera systems, as shown in Fig. 4.3 [7]. The aim is to produce high-quality content for virtual reality applications and interactive film formats.

To capture 3D video objects, the first step is to arrange a number of synchronized high-resolution camera pairs that capture the scene from several directions. The small baseline within each stereo pair avoids any masking and view dependent surface reflections, thus enabling the estimation of robust depth maps. In addition, the usage of several stereo pairs, enables global object capture via fusion of the individually recorded object surfaces.

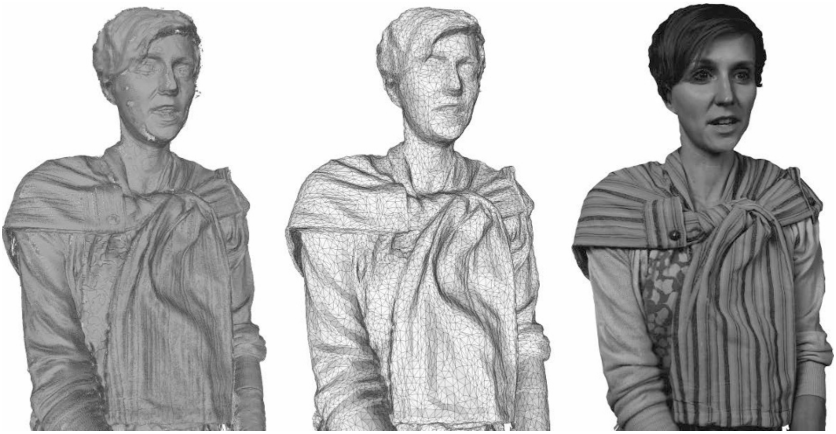


Fig 4.3 3D person reconstruction; *left:* 3D geometry as originally reconstructed point cloud; *center:* 3D geometry as reduced wireframe; *right:* textured model with projected videos from the multi-camera system, source [7] (Fraunhofer HHI)

Depth maps are estimated from the stereo pairs using a patch-based matching approach. Here, the dependencies are minimized in such a way that the matching between corresponding patches can be evaluated independently, permitting efficient GPU parallelization and real-time implementation [53]. Propagating depth information from neighboring and temporally preceding patches ensures spatial and temporal smoothness and improved robustness. Lighting and color compensation reduces differences between the images from different cameras. Depending on the degrees of freedom of the patch used, the accuracy of the depth estimation is much below 1 mm, as shown in Fig. 4.4. In the next stage, the point clouds for all stereo pairs are fused, and points with incorrectly estimated 3D positions are eliminated via consistency tests across the individual point clouds. As a result, a 3D point cloud of the entire 3D video object is produced with approx. 50 million points (see Fig. 4.3 *left*).

Using the color information, a dynamic 3D object representation can be produced at this stage, however only as a colored point cloud without a (closed) surface. Hence, in the next step, the point cloud is triangulated, i.e. all points are connected by triangles. For consistent triangulation, Poisson surface reconstruction is used [19], placing a smooth surface through the point clouds and normals. Then, the triangulated 3D model can be simplified, e.g. via successive point elimination in smooth surface regions. The aim here is to retain surface details while reducing the

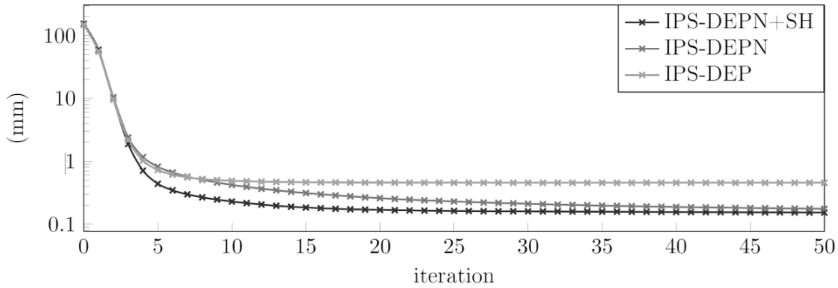


Fig. 4.4 Accuracy of iterative depth estimation, in comparison to known reference objects (Fraunhofer HHI)

number of 3D points and triangles at the same time. The result is a compact wireframe as shown in Fig. 4.3, *center*. Here, the original 50 million points were reduced to approximately 10,000 points [7].

In order to guarantee a high quality of texture and color of the 3D model, the original video data from the cameras is now projected onto the wireframe. To achieve the highest detail information, the most suitable camera view in terms of spatial resolution is identified for each surface triangle of the 3D model. Then, color information is primarily projected from that camera. Interpolating between neighboring triangles avoids visible breaks and inconsistencies in the surface texture. The resulting textured object is depicted in Fig. 4.3 *right*.

This method was developed by Fraunhofer HHI as a fully-automated method for producing highly realistic, dynamic 3D video objects and integrating them into virtual worlds [7]. These objects are characterized in particular by their natural appearance and realistic movements, thus representing a cornerstone of interactive 3D video worlds.

When integrating the 3D video objects into new virtual environments, they are displayed in the same way as they were captured during the recording method. Often, the lighting, body pose or motion require customization in order to tailor them to the environment or to enable interaction with the video objects. To do this, semantic body models (avatars) are customized to the individual 3D geometry and motion patterns are learnt from measured data [9]. Using the skeleton assigned to the model, corrections to the body pose can be carried out that are then translated to the individual geometry, e.g. to represent interactions in the virtual world. Moreover, individual motion sequences can be recombined and seamlessly superimposed [34] in order to realize complex representations or new interactive forms of media viewing.

4.5 Summary and Outlook

In this chapter, we have shown the development of digital video data towards a globally dominant data form, and explained the development of video processing. Two of the most successful technologies, with Fraunhofer HHI strongly involved in the development, are video coding and production of three-dimensional dynamic video objects. Both fields will continue developing, leading to common topics for research and production, and shape a number of fields in digital transformation.

In video coding, a successor standard will be developed in the coming years that once again provides improved compression for video data. This is particularly necessary as in the process towards higher video resolutions from HD via UHD/4K, 8K has already been announced and will once again multiply the uncoded video data rate. In the case of 3D video object reconstruction, further improvements in 3D display technology together with fast, high-quality production of natural 3D video scenes will allow for interactive 3D video worlds to be created where reality and virtuality will merge seamlessly. The successful distribution of 3D video scenes also requires standardized formats and compression processes that will be developed in the coming years. The first steps have already been taken with extensions to depth-based 3D video coding and standardization plans for omnidirectional 360° video within the new coding standards. Nevertheless, further research is required, in order to internationally standardize methods for efficient, dynamic 3D scene coding and thus to establish globally successful systems and services for interactive 3D video worlds.

Sources and literature

[1] *Advanced Video Coding,* Rec. ITU-T H.264 and ISO/IEC 14496-10, Oct. 2014.

[2] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform", IEEE Transactions on Computers, vol. C-23, issue 1, pp. 90–93, Jan. 1974

[3] D. Blumenthal-Barby, P. Eisert, "High-Resolution Depth For Binocular Image-Based Modelling", Computers & Graphics, vol. 39, pp. 89-100, Apr. 2014.

[4] J. Boyce, Y. Yan, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding (HEVC) Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, issue 1, pp. 20-34, Jan. 2016

[5] Cisco Systems, Inc. Cisco visual networking index, "Forecast and methodology, 2016-2021", White paper, June 2017. Retrieved June 8, 2017, from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf .

[6] *Coding of audio-visual objects – Part 2: Visual,* ISO/IEC 14496-2, 2001

[7] T. Ebner, I. Feldmann, S. Renault, O. Schreer, and P. Eisert, "Multi-view reconstruction of dynamic real-world objects and their integration in augmented and virtual reality applications", Journal of the Society for Information Display, vol. 25, no. 3, pp. 151–157, Mar. 2017, doi:10.1002/jsid.538

[8] H. Everett III, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources", Operations Research, vol. 11, issue 3, pp. 399–417, June 1963

[9] P. Fechteler, A. Hilsmann, P. Eisert, Example-based Body Model Optimization and Skinning, Proc. Eurographics, Lisbon, Portugal, May 2016

[10] D. Flynn, D. Marpe, M. Naccari, T. Nguyen, C. Rosewarne, K. Sharman, J. Sole, and J. Xu, "Overview of the range extensions for the HEVC standard: Tools, profiles and performance", IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, issue 1, pp. 4–19, Jan. 2016

[11] E. François, C. Fogg, Y. He, X. Li, A. Luthra, and A. Segall, "High dynamic range and wide color gamut video coding in HEVC: Status and potential future enhancements", IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, issue 1, pp. 63–75, Jan. 2016

[12] C.-M. Fu, E. Alshina, A. Alshin, Y.-W. Huang, C.-Y. Chen, C.-Y. Tsai, C.-W. Hsu, S. Lei, J.-H. Park, and W.-J. Han, "Sample adaptive offset in the HEVC standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, issue 12, pp. 755–1764, Dec. 2012

[13] J. Furch, A. Hilsmann, P. Eisert, "Surface Tracking Assessment and Interaction in Texture Space", Computational Visual Media, 2017. Doi: 10.1007/s41095-017-0089-1

[14] Generic coding of moving pictures and associated audio information – Part 2: Video, Rec. ITU-T H.262 and ISO/IEC 13818-2, Jul. 1995.

[15] P. Helle, S. Oudin, B. Bross, D. Marpe, M. O. Bici, K. Ugur, J. Jung, G. Clare, and T. Wiegand, "Block Merging for Quadtree-Based Partitioning in HEVC", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, issue 12, pp. 1720-1731, Dec. 2012.

[16] C. Hellge, E. G. Torre, D. G.-Barquero, T. Schierl, and T. Wiegand, "Efficient HDTV and 3DTV services over DVB-T2 using Multiple PLPs with Layered Media", IEEE Communications Magazine, vol. 51, no. 10, pp. 76-82, Oct. 2013

[17] *High Efficiency Video Coding,* Rec. ITU-T H.265 and ISO/IEC 23008-2, Oct. 2014.

[18] A. Hilsmann, P. Fechteler, P. Eisert, "Pose Space Image-based Rendering", Computer Graphics Forum (Proc. Eurographics 2013), vol. 32, no. 2, pp. 265-274, May 2013

[19] M. Kazhdan, M. Bolitho, H. Hoppe, "Poisson surface reconstruction", Symposium on Geometry Processing, 61-70, 2006.

[20] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, issue 12, pp 1792–1801, Dec. 2012

[21] D. Marpe, H. Schwarz, S. Bosse, B. Bross, P. Helle, T. Hinz, H. Kirchhoffer, H. Lakshman, T. Nguyen, S. Oudin, M. Siekmann, K. Sühring, M. Winken, and T. Wiegand, "Video Compression Using Nested Quadtree Structures, Leaf Merging, and Improved Techniques for Motion Representation and Entropy Coding", IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, issue 12, pp. 1676-1687, Dec. 2010, Invited Paper.

[22] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, issue 7, pp. 620–636, July 2003

[23] D. Marpe, T. Wiegand and G. J. Sullivan, "The H.264/MPEG4 Advanced Video Coding Standard and its Applications", IEEE Image Communications Magazine, vol. 44, is. 8, pp. 134-143, Aug. 2006.

[24] T. Matsuyama, X. Wu, T. Takai, T. Wada, "Real-Time Dynamic 3-D Object Shape Reconstruction and High-Fidelity Texture Mapping for 3-D Video", Transaction on Circuits and Systems for Video Technology, vol. 14, issue 3, pp. 357-369, March 2004.

[25] P. Merkle, K. Müller, D. Marpe and T. Wiegand, "Depth Intra Coding for 3D Video based on Geometric Primitives", IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 3, pp. 570-582, Mar. 2016.

[26] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient Prediction Structures for Multiview Video Coding", invited paper, IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1461-1473, Nov. 2007.

[27] K. Müller, P. Merkle, and T. Wiegand, "3D Video Representation Using Depth Maps", Proceedings of the IEEE, Special Issue on 3D Media and Displays, vol. 99, no. 4, pp. 643 – 656, April 2011.

[28] K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, H. Rhee, G. Tech, M. Winken, and T. Wiegand: "3D High Efficiency Video Coding for Multi-View Video and Depth Data", IEEE Transactions on Image Processing, Special Section on 3D Video Representation, Compression and Rendering, vol. 22, no. 9, pp. 3366-3378, Sept. 2013.

[29] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View Synthesis for Advanced 3D Video Systems", EURASIP Journal on Image and Video Processing, Special Issue on 3D Image and Video Processing, vol. 2008, Article ID 438148, 11 pages, 2008. doi:10.1155/2008/438148.

[30] T. Nguyen, P. Helle, M. Winken, B. Bross, D. Marpe, H. Schwarz, and T. Wiegand, "Transform Coding Techniques in HEVC", IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 6, pp. 978-989, Dec. 2013.

[31] A. Norkin, G. Bjøntegaard, A. Fuldseth, M. Narroschke, M. Ikeda, K. Andersson, M. Zhou, and G. Van der Auwera, "HEVC deblocking filter", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, issue 12, pp. 1746–1754, Dec. 2012

[32] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards – Including High Efficiency Video Coding (HEVC)", IEEE Transactions on Circuits and Systems for Video Technology, Dec. 2012.

[33] G. A. Østerberg, "Topography of the layer of rods and cones in the human retina", Acta Ophthalmologica, vol. 13, Supplement 6, pp. 1–97, 1935

[34] W. Paier, M. Kettern, A. Hilsmann, P. Eisert, "Video-based Facial Re-Animation", Proc. European Conference on Visual Media Production (CVMP), London, UK, Nov. 2015.

[35] J. Posada, C. Toro, I. Barandiaran, D. Oyarzun, D. Stricker, R. de Amicis, E. Pinto, P. Eisert, J. Döllner, and I. Vallarino, "Visual Computing as a Key Enabling Technology for Industrie 4.0 and Industrial Internet", IEEE Computer Graphics and Applications, vol. 35, no. 2, pp. 26-40, April 2015

[36] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders", IEEE Transactions on Image Processing, vol. 3, issue 5, pp. 533–545, Sept. 1994

[37] F. Rocca and S. Zanoletti, "Bandwidth reduction via movement compensation on a model of the random video process", IEEE Transactions on Communications, vol. 20, issue 5, pp. 960–965, Oct. 1972

[38] A. Saxena and F. C. Fernandes, "DCT/DST-based transform coding for intra prediction in image/video coding", IEEE Transactions on Image Processing, vol. 22, issue 10, pp. 3974–3981, Oct. 2013

[39] T. Schierl, K. Grüneberg, and T. Wiegand, "Scalable Video Coding over RTP and MPEG- 2 Transport Stream in Broadcast and IPTV Channels", IEEE Wireless Communications Magazine, vol. 16, no. 5, pp. 64-71, Oct. 2009.

[40] T. Schierl, T. Stockhammer, and T. Wiegand, "Mobile Video Transmission Using Scalable Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Scalable Video Coding, vol. 17, no. 9, pp. 1204-1217, Sept. 2007, Invited Paper

[41] D. Schneider, A. Hilsmann, P. Eisert, "Warp-based Motion Compensation for Endoscopic Kymography", Proc. Eurographics, Llandudno, UK, pp. 48-49, Apr. 2011.

[42] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard", IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Scalable Video Coding, vol. 17, no. 9, pp. 1103-1120, Sept. 2007, Invited Paper

[43] R. Sjöberg, Y. Chen, A. Fujibayashi, M. M. Hannuksela, J. Samuelsson, T. K. Tan, Y.-K. Wang, and S. Wenger, "Overview of HEVC High-Level Syntax and Reference Picture Management", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1858–1870, Dec. 2012.

[44] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Müller, and M. Lang, "3D Video Post-Production and Processing", Proceedings of the IEEE (PIEEE), Special Issue on 3D Media and Displays, vol. 99, issue 4, pp. 607-625, April 2011

[45] T. Stockhammer, M. M. Hannuksela, and T. Wiegand, "H.264/AVC in wireless environments", IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 657–673, July 2003.

[46] G. J. Sullivan and R. L. Baker, "Efficient quadtree coding of images and video", IEEE Transactions on Image Processing, vol. 3, issue 3, pp. 327–331, May 1994

[47] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[48] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression", IEEE Signal Processing Magazine, vol. 15, no. 6, pp. 74–90, June 1998

[49] V. Sze and M. Budagavi, "High throughput CABAC entropy coding in HEVC", IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, issue 12, pp. 1778–1791, Dec. 2012

[50] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro and Y. K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on HEVC Extensions and Efficient HEVC Implementations, vol. 26, no. 1, pp. 35-49, Jan. 2016.

[51] K. Ugur, A. Alshin, E. Alshina, F. Bossen, W.-J. Han, J.-H. Park, and J. Lainema, "Mo- tion compensated prediction and interpolation filter design in H.265/HEVC", IEEE Journal of Selected Topics in Signal Processing, vol. 7, issue 6, pp. 946–956, Dec. 2013.

[52] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/AVC Standard", Proceedings of the IEEE, Special Issue on "3D Media and Displays", vol. 99, issue 4, pp. 626-642, April 2011, Invited Paper.

[53] W. Waizenegger, I. Feldmann, O. Schreer, P. Kauff, P. Eisert, "Real-Time 3D Body Reconstruction for Immersive TV", Proc. IEEE International Conference on Image Processing (ICIP), Phoenix, USA, Sep. 2016.

[54] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control", In Proc. of International Conference on Image Processing (ICIP), vol. 3, pp. 542–545, 2001.

[55] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra, "Rate- distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, issue 2, pp.182–190, April 1996

[56] T. Wiegand and H. Schwarz, "Source Coding: Part I of Fundamentals of Source and Video Coding", Foundations and Trends in Signal Processing, vol. 4, no. 1-2, pp. 1-222, Jan. 2011, doi:10.1561/2000000010

[57] T. Wiegand and H. Schwarz, "Video Coding: Part II of Fundamentals of Source and Video Coding", Foundations and Trends in Signal Processing, vol. 10, no. 1-3, pp 1-346, Dec. 2016, doi:10.1561/2000000078

[58] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards", IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, issue 7, pp. 688–703, July 2003

[59] T. Wiegand and G. J. Sullivan, "The H.264/AVC Video Coding Standard {Standards in a nutshell}", IEEE Signal Processing Magazine, vol. 24, no. 2, March 2007.

[60] *Video codec for audiovisual services at p × 64 kbits,* Rec. ITU-T H.261, Nov. 1988.

[61] *Video coding for low bit rate communication.* Rec. ITU-T H.263,Mar. 1996.

Audio Codecs

5

Listening pleasure from the digital world

Prof. Dr. Karlheinz Brandenburg · Christoph Sladeczek
Fraunhofer Institute for Digital Media Technology IDMT

Summary

The development of music recording and reproduction has been characterized by the quest for perfection ever since its inception under Thomas Alva Edison. This is true for all fields of recording and reproduction technology, from microphones and sound storage media to loudspeaker technology. Concert-quality sound reproduction which matches the original with complete fidelity remains the goal of research. This can only be achieved, however, if all of the factors involved in listening are addressed, whether they are acoustic, psychoacoustic, or psychological. Fraunhofer IDMT's further development of wave field synthesis technology as a marketable product – already in use in highly demanding open air and opera house productions – plays a key role here. The uncomplicated synchronous storage and transmission of metadata provided by the current MPEG-H 3D Audio coding method allows listeners at home to interactively shape their listening experience.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_5

5.1 Introduction: The dream of high fidelity

Even Thomas Alva Edison already dreamt of perfect sound reproduction. Agents marketing his phonograph as a consumer product travelled the world, conducting what were perhaps the earliest sound tests: Audiences listened amazed as music was performed in darkened halls – first as a live performance by a singer or cellist, for example, and then as a phonograph recording of the same piece. Many listeners found the recording's sound quality so good that they could not tell the difference [8]. From this we can conclude that our assessment of sound quality is particularly related to our expectations of a medium. Background noises such as hisses or crackles were not part of the music and were thus not heard.

Ever since then, research on how to deliver the most perfect sound reproduction possible has been ongoing. *High fidelity* has been the common term used to describe this for many years now. If we were to assess Edison's test by modern standards, there would be little in common between the phonograph's wax cylinder and our current understanding of hi-fi, even though considerable sound quality was already achieved in those days. Today, a good reproduction setup could probably pass the test for individual musical instruments, but not for larger ensembles such as a string quartet or a symphony orchestra. Sound reproduction that creates a perfect illusion within the room still remains just out of reach. In recent decades, however, significant progress has been made in terms of both loudspeaker and headphone playback. The goal of complete immersion – diving into an alien sound environment – is thus getting ever closer. Keywords here are stereophonics, surround sound, wave field synthesis, ambisonics, and object-based audio signal storage.

5.2 Hi-fi technologies from analog to digital

In the early years of hi-fi technology, the research emphasis lay on the different elements of the processing chain: microphones, analog recording processes (tapes, vinyl records), playback devices for tapes and records, amplifiers, loudspeakers and headphones. At that, analog audio recording processes such as tape recorders have fundamental limitations at the level of sound quality they are able to provide. Ever greater progress has been made over the decades in optimizing the components: whereas the tube technology amplifiers display nonlinearity of transmission (output voltage over input voltage) – audible even to less practiced listeners – modern amplifier technology (implemented in analog integrated circuits or as so-called class-D amplifiers using mixed analog/digital circuit technology) is close to perfect. Microphones, too, are near to physical limits for professional applications. Loudspeakers and headphones still remain the weakest link in this transmission chain. If our hearing were not as good as it is at adapting to the deviations from the ideal frequency response curve, a lot of electronically amplified music would sound off-key and alien.

The various music recording media represented key milestones in the development of hi-fi technology: the phonograph cylinder, gramophone records, vinyl long-play records, reel-to-reel tape recorders, compact cassettes and compact discs. Over time, ever-newer technologies had to be introduced in order to provide a higher potential sound quality, increased usefulness (for example, so people could also make their own recordings on tape recorders) and, in particular, convenient storage medium handling. The first digital music storage medium to gain widespread acceptance was the CD. Its parameters (16-bit resolution, 44,100 Hz sampling frequency, two sound channels) were the state of the art at the time and were designed, according to detailed tests, to facilitate "perfect" sound reproduction.

The underlying optical storage medium – the CD-ROM – became the storage medium for software distribution for several years. As a storage medium for audio-visual applications such as short films, the CD-ROM was also one of the first innovations to initiate the development of processes for video and audio coding in the Moving Pictures Experts Group (MPEG, officially ISO/IEC JTC1 SC29 WG11). The standards produced by this standardization committee simultaneously represented the state of the art in their respective fields when they were adopted. From MPEG-1 Audio Layer-3 (known as MP3) through MPEG-2/4 Advanced Audio Coding (AAC) and further developments such as HE-AAC to today's MPEG-H, this family of audio coding methods is now incorporated into around 10 billion devices. Each new generation of processes brought improved coding efficiency (identical sound quality at lower bit rates/better sound quality at equally low bit rates) as well as increased flexibility (for example, supporting surround sound processes). Through all these years, Fraunhofer IIS has made significant contributions to these processes and their standardization.

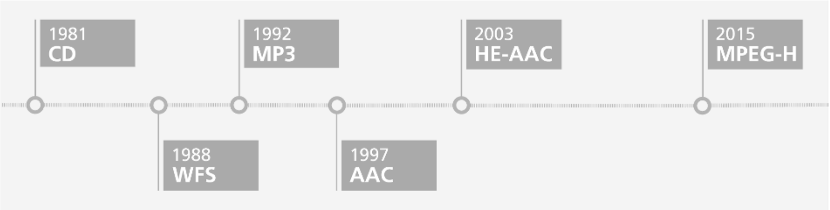


Fig. 5.1 Selective listing of important developments in digital audio technology in recent decades (Fraunhofer IDMT)

Whereas for a long time the standard for home music listening remained two-channel stereo, cinema technology saw the introduction of surround sound, designed to facilitate an immersive sound experience. Since MPEG-2, the audio coding methods standardized by MPEG have also supported 5.1-channel surround sound as a minimum. Newer methods, especially for professional applications, will be described in later sections. If we compare the sound experience of today with the dreams of Edison's days then we have to accept that the perfect audio illusion has still not been achieved. This is not due to the data reduction processes but to the fundamental technology of sound reproduction. Today, the reproduction of material in mono or in two-channel stereo is possible to such a perfect degree – even when using AAC data reduction, for example – that blind tests show that it is impossible to distinguish it from a reference recording. In short, the Edison blind test is successful as long as there is only one instrument or musician on stage. Reproducing the sound experience of listening to a symphony orchestra, however, remains a challenge.

5.3 Current research focus areas

5.3.1 The ear and the brain

Recent developments in hi-fi technologies are closely related to the scientific discipline of psychoacoustics. The aim of psychoacoustic research is to discover the relationships between stimuli – i.e. sounds such as music as well as all kinds of other audio signals – and human perception. Two areas are particularly key to human hearing:

- The ear (including the outer ear and pinna, the auditory canal, the eardrum, the middle ear, and the inner ear with the cochlea and sensory hairs)
- The brain (by processing the electrochemical reactions that are triggered by the motion of the sensory hairs connected to the neurons – the so-called "firing" of neurons), including the higher region of our brain, responsible for complex perception and the identification of sound.

To simplify, we can say that the ear's characteristics define the general parameters of our hearing such as the frequency range of discernable sounds and the dynamic range, from the quietest audible sounds through to sounds that can lead to damage of the inner ear. The everyday phenomenon of masking – that is, the obscuring of quieter sounds by louder ones – is a significant effect for audio coding in particular, and can be explained by the mechanics of the inner ear. Sounds of one frequency mask quieter tones of the same or neighboring frequencies. This effect is utilized in audio coding (MP3, AAC, etc.), where frequencies with reduced accuracy are transmitted in such a way that the differential signals are masked by the actual music signal and thus no difference to the input signal is audible.

Since these masking effects are mechanically produced– and since the stimuli transmitted from the ear to the brain are already "data reduced", so to speak – they are stable. Even with extensive listening training, people are unable, during blind tests, to distinguish pieces of music coded at a sufficiently high bit rate (using AAC for example) from the original.

Nevertheless, the processing that happens in the brain is utterly vital to the auditory impression. Research into the so-called cognitive effects of hearing has been ongoing for decades, but we have to admit that our understanding of these processes is significantly more limited than our understanding of the workings of the ear. In short, feedback effects from the higher levels of the brain, and especially expectations of the sound play a role. All spatial hearing is connected with these complex mechanisms: the conjunction of signals from both ears (binaural spatial hearing) takes place in the brain. Here, differences in time and especially phase of the sound heard by both ears are evaluated.

A further key effect is the filtering of signals by the pinna and head. These effects are usually described as the outer ear transfer function (technically, *HRTF* or *Head Related Transfer Function*). This function, which varies from person to person, is something we will keep in mind as we continue through this chapter. It is this function that enables us to roughly ascertain a sound's direction of origin even when using one only ear. As things currently stand, spatial hearing in hi-fi technology is determined by how "coherent" our perception of a sound is, given where we expect the sound to come from (especially when we can see its source) and given the playback system and any additional effects. The greater the divergence in our perception, the more frequently localization errors occur and – especially in the case of headphone playback – the more confusion between sound coming from behind or the front or sound source localization within the head (instead of outside the head) is observed. All of these effects are highly individual, varying significantly over time. The brain can be trained to perceive certain illusions more frequently.

Research is currently being carried out into precisely these issues:

- How does the listening environment – the room and the reflections of the sound from the walls and furniture – affect the acoustic illusion that hi-fi technology tries to create?
- How much are we influenced by our expectations, and to what extent do we fail to perceive changes in the sound even though they are clearly measurable? On

the other hand, to what degree do we hear differences that are purely based on our expectations?

These questions are closely related to an apparent contradiction that is especially identified in listening tests which use high quality playback equipment and storage processes: the closer a playback system is placed physically to the original signal, the sooner psychological factors (cognitive effects in the brain) lead to us being certain we are hearing something that subsequently disappears in the blind test, when statistically analyzed. This is the reason why many hi-fi enthusiasts insist on special cabling for connecting their stereo equipment to their domestic power supply, while others will only accept equipment and formats that match the hearing range of bats and dogs (ultrasound) but are irrelevant to human perception. The widespread aversion to audio coding methods can also be traced back to these psychological effects.

But these observations still do not mean that our task of creating the perfect sound illusion is completed, even in terms of spatial sound. Our brain perceives the surrounding space and the distribution of sound waves with a high degree of precision, especially in terms of the temporal sequence of various audio signals and their reflections in the room. Research into the creation of this illusion is currently being carried out in several places, particularly at Fraunhofer IDMT in Ilmenau. Perfect sound in the room is an old dream, but one that is increasingly being fulfilled. Methods for spatial sound reproduction via a larger number of loudspeakers are helpful here, as we will discuss in the following sections.

5.3.2 From audio channels to audio objects

In classical audio production, loudspeaker signals are saved to the sound storage medium as a result of mixing. Instruments and sound sources are distributed spatially in the audio signal by means of different volume weighting of the loudspeakers. This process is known as *panning* and is set separately for each sound source by the sound engineer so that voices, for example, are heard by the user in the middle of the stereo panorama during playback, and instruments such as guitars and drums are heard to the left and right. Specific guidelines need to be followed in order to correctly perceive the spatial mix intended by the sound engineer. The stereo speakers, for example, must be set up in the same positions as they are in the recording studio, and the listening location must form an equilateral *stereo triangle* with the speakers. The same holds for stereo playback processes such as 5.1 surround sound, for example, for which supplementary speakers are added. Since the spatial sound reproduction in each case is only correct for a single listening location, this location is known as the *sweet spot*. If we look at actual playback locations such as cinemas, theatres, and the home, for example, we see that for most listeners the ideal playback location cannot be maintained.

Ever since the initial beginnings of the development of spatial sound playback processes, the desire has been to record and reproduce the sound field of a virtual sound source in such a way as to provide the correct spatial sound impression for all listeners. One attempt uses loudspeaker arrays to synthesize a sound field in a way that is physically correct. To achieve this goal, Steinberg and Snow in 1934 published the principle of the *acoustic curtain* [11], where a fixed network of microphones in the recording studio connected to loudspeakers in the playback room, is used to record and reproduce the sound field of a source. Steinberg and Snow were using Huygens ' Principle here, which states that an elementary wave can be produced by the superimposition of many individual secondary sound sources.

At the time, however, it was still not technologically possible to implement this complex arrangement in practice, which was why Steinberg and Snow limited their system to three loudspeakers. The use of these three loudspeakers represents the beginnings of the stereophonic recording and playback technology that was extended with additional channels in subsequent years.

In the 1980s, Guus Berkhout carried out research into acoustic holography procedures for use in seismology. In doing so, he used arrays of microphones that recorded the reflection patterns of special sound signals from different layers of the earth, thus providing information about the substances contained in the ground. Because of his personal interest in acoustics in general, he suggested that the technology he had developed could be reversed and used for loudspeaker arrays. This marked the beginnings of the development of wave field synthesis technology, where arrays of loudspeakers are used to synthesize – within specific limits – a physically correct sound field for a virtual sound source [1]. The underlying principle is illustrated in Fig. 5.2.

In the years that followed, the technology was developed at TU Delft to the point where it could be presented as a functional laboratory demonstrator in 1997 [12]. A key characteristic of wave field synthesis technology is its object-based approach. In contrast to classical sound reproduction processes, audio objects are stored rather than loudspeaker signals as a result of the spatial sound mixing. An audio object is defined as a (mono) signal that contains audio information – for example a violin or a female voice – together with its associated metadata, which describes properties such as the position, volume or type of audio object. In order to investigate this new technology and its associated production, storage, transmission and interaction requirements with respect to a potential introduction to the market, a consortium was

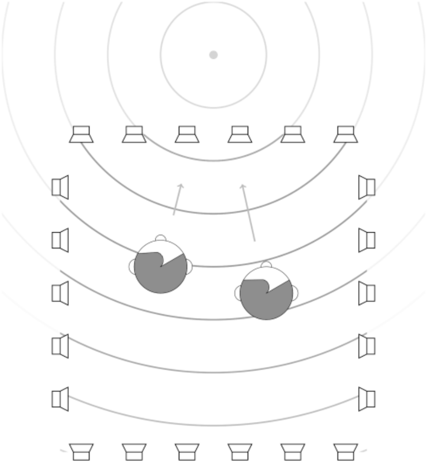


Fig 5.2 Schematic representation of wave field synthesis. An array of loudspeakers surrounding the listening space is controlled such that a physically correct sound field of a virtual sound source is produced by superimposing the individual loudspeaker outputs. (Fraunhofer IDMT)

formed in 2001, consisting of industry and research and development, under the banner of the EU CARROUSO project [2]. As a key outcome of this project Fraunhofer IDMT was able to present a first marketable product prototype installation at the Ilmenau cinema in 2003.

5.3.3 Audio objects in practice

In contrast to channel-based sound reproduction, where fully mixed loudspeaker signals are simply played back, with object-based sound reproduction the signals have to be calculated interactively. This concept is illustrated in Fig. 5.3. An object-based reproduction system, at its core, consists of an audio renderer that produces the loudspeaker signals [6]. To do this, the coordinates of the loudspeakers must be known to the renderer. Based on this information, the metadata from the audio objects and the corresponding audio signals are streamed to the renderer in real time. In this process, the system is completely interactive such that each audio object can be positioned at will. An additional distinguishing characteristic of the object-based approach is that the audio renderer is able to make use of various playback technologies. Instead of using a wave field synthesis-based renderer for producing the loudspeaker signals, a system based on binaural technology, for example, can be used. It is thus also possible for the spatial audio mix produced for loudspeaker playback, for example, to provide a plausible three-dimensional sound

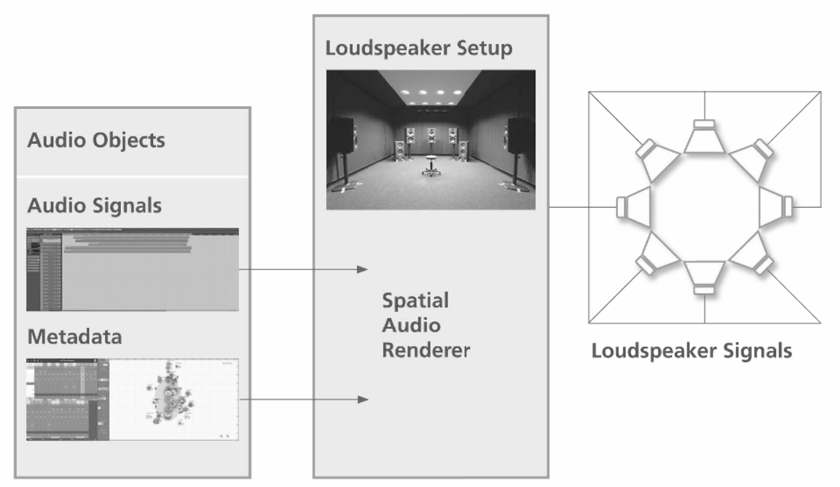


Fig. 5.3 Object-based sound reproduction concept. Based on an object-based description of the spatial acoustic scene together with the input signals, the spatial audio renderer produces the output signals, knowing the loudspeaker coordinates. (Fraunhofer IDMT)

perception via headphones [7]. This is far more difficult with a channel-based audio approach.

One additional interesting distinguishing feature of the object-based audio playback concept is sideways compatibility with channel-based audio content [3]. Here, the loudspeaker signals are reproduced via audio objects so that a virtual loudspeaker setup is produced. In this way, channel-based audio content is practically independent of the actual number of playback speakers. In addition, the user has the option of changing the virtual loudspeaker setup and can thus vary parameters such as stereo width or spatial envelopment intuitively.

For audio objects to be able to be used intuitively in a practical application, the technology must be capable of being seamlessly incorporated into an audio environment that is familiar to the sound engineer [9]. To do this, it is necessary that the technology mirrors established interfaces and interaction paradigms. Fig. 5.4 shows one potential possibility for integration.

The object-based sound reproduction technology is thus based on a rendering PC that is equipped with a professional sound card with standard multi-channel audio transmission formats. A sound source of the user's choice can then be connected via this interface. This is shown in the illustration by a digital audio workstation (DAW) and a mixer symbol. For the audio objects to be able to be positioned intuitively a

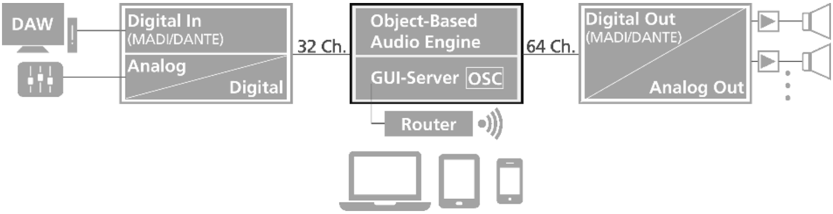


Fig 5.4 Sample integration of object-based sound reproduction into existing professional audio production structures (Fraunhofer IDMT)

graphical user interface (GUI) is required; in the example above, this is provided by a web-based application accessed via a web browser. In this case, the server providing the user interface also runs on a rendering PC. An example of this kind of user interface is shown in Fig. 5.5.

The interface here is divided into two regions. On the right-hand side is the positioning region where the audio objects can be positioned. Here, audio objects are represented by round icons. In order to provide the sound engineer with a familiar interface, the audio objects' properties are listed clearly on the interface's left-hand side in the form of channel strips. The sound engineer is thus able to shape the playback of the individual audio objects in the same way as at a mixing desk.

Fig. 5.5 Graphical user interface for object-based audio production (Fraunhofer IDMT)

SpatialSound Wave for professional sound reinforcement

The work undertaken at Fraunhofer IDMT to develop wave field synthesis technology towards a marketable product has been finding application in various fields for a number of years now. Especially for live performances, the trend in recent years has been to boost the sound experience by using spatial sound reinforcement. One example of this is the Bregenz Festival, where wave field synthesis technology has been used since 2005 to create artificial concert hall acoustics in an open-air setting (for the approx. 7,000-seater main auditorium) where this would otherwise be lacking. To do this, the entire seating area was surrounded by a line of nearly 800 loudspeakers. The installation is shown in Fig. 5.6.

Since a "real" wave field synthesis installation would require a very large number of loudspeakers and amplifier channels, its use is limited by the high hardware costs. In addition, while for certain installations this kind of system would be very desirable, building constraints make it impossible. Theatres and opera houses represent this kind of case since they are often protected buildings where the visual appearance of the performance hall is not permitted to be changed. At the same time, however, the need often arises – especially in the case of innovative productions – to use acoustics to involve the audience more fully in the action. Since hearing is the only one of our senses that is active in all spatial directions, this effect can only be achieved if the playback system allows for sound

Fig. 5.6 Object-based sound reproduction for the Bregenz Festival live sound reinforcement (Fraunhofer IDMT)

Fig. 5.7 Object-based audio reproduction in the Zurich Opera House (Dominic Büttner/Zurich Opera House)

reinforcement from all spatial directions. Often, opera houses and theatres are equipped with loudspeaker installations of 80 speakers and more, but these are distributed loosely and in three dimensions throughout the space. In order to support these precise installations in terms of acoustics, the wave field synthesis algorithm was altered so that, taking account of human perception, audio objects can still be stably localized over a large listening area, but allowing for greater distance and three-dimensional distribution of speakers. This *SpatialSound Wave* technology has been applied in several prestigious venues such as the Zurich Opera House shown in Fig. 5.7.

Alongside live sound reinforcement, there is an additional area of application in the acoustic support of large-screen playback systems. Planetariums are good examples of dome projection installations that, in recent years, have moved away from classical visualizations of the night sky towards being entertainment adventure venues. Although dome projection offers an impressive, enveloping image from different seating positions, installations historically often only featured a few loudspeakers using channel-based playback formats. That meant that in practice, the image was spread across the dome, but the sound came from individual loudspeakers below the projection surface. Plausible reproduction, however, requires images and sound objects to be coordinated spatially, in the same way that SpatialSound Wave technology does, by allowing positioning of the loudspeakers behind the projection surface. Fig. 5.8 shows a typical SpatialSound Wave technology application in the world's oldest functioning planetarium, the Zeiss Planetarium in Jena.

Fig. 5.8 Object-based audio reproduction for large-scale projection systems in-room sound reinforcement (Fraunhofer IDMT)

MPEG-H 3D Audio

Object-based audio productions require new file formats for storage and transmission since, along with pure audio signals, metadata also needs to be reproduced synchronously. The MPEG-4 standard already supported storing audio objects via the so-called BIFS (Binary Format for Scenes) scene description, but this was hardly ever used due to the great complexity of the MPEG-4 standard [5].

MPEG-H is a new standard, which enables both the transmission and storage of high-quality video and audio. By means of the relevant component for audio data – known as MPEG-H 3D Audio – it is now possible to store the most diverse audio formats in a standardized structure. Alongside the ability to save channel-based audio unambiguously even with a large number of speaker channels, this innovation offers storage of audio objects and higher-order ambisonics signals. By supporting various audio formats it becomes apparent that the standard will become accepted over the coming years as the container format for all new immersive audio playback methods.

The standard is currently being introduced in the field of broadcasting. Alongside the loudspeaker system scaling advantages already mentioned, audio objects offer new possibilities for interaction here. One example is in the transmission of sporting events where today a sound engineer mixes commentator vocals and stadium atmosphere sound in the broadcasting vehicle, so that they can then be transmitted to the end user in stereo or surround sound format. Depending on what end device the consumer is using, problems can arise with the intelligibility of speech, for example, if the commentator's voice is obscured by the atmosphere in the stadium. Since signal elements in a stereo signal cannot be changed retrospectively anymore (or can only be changed very little), one potential solution is the transmission as audio objects. This way, the user has the option, for example, to change the volume of the commentator at home to restore speech intelligibility or to concentrate completely on the atmosphere in the stadium [10].

5.4 Outlook

It has been a long journey from Edison's "sound tests" to current sound reproduction, assisted by complex digital signal processing algorithms. Modern technology enables us to listen to music comfortably and at near perfect sound quality whether we are at home or on the road. Certain problems, such as achieving the perfect illusion, have however still only been partially solved. But as fundamental research continues and is applied to the latest audio signal storage and playback standards, both live and pre-recorded music will, over the coming decades, deliver even more listening pleasure.

Sources and literature

[1] Berkhout A.W,: „A Holographic Approach to Acoustic Control", JAES, Volume 36, Issue 12, pp. 977-995, December 1988.

[2] Brix S., Sporer T., Plogsties J., „CARROUSO – An European Approach to 3D Audio", 110th Convention of the AES, Amsterdam, The Netherlands, May 2001.

[3] Brix S., Sladeczek C., Franck A., Zhykhar A., Clausen C., Gleim P.; „Wave Field Synthesis Based Concept Car for High-Quality Automotive Sound", 48th AES Conference: Automotive Audio, Hohenkammern, Germany, September 2012.

[4] Brandenburg K., Brix S., Sporer T.: „Wave Field Synthesis – From Research to Applications", European Signal Processing Conference (EUSIPCO) 2004, September, Vienna, Austria.

[5] Herre J., Hilpert J., Kuntz A., Plogsties J.: „MPEG-H 3D Audio – The New Standard for Coding of Immersive Spatial Audio", IEEE Journal of Selected Topics in Signal Processing, Volume 9, No. 5, pp. 770-779, August 2015.

[6] Lembke S., Sladeczek C., Richter F., Heinl, T., Fischer C., Degardin P.; „Experimenting with an Object-Based Sound Reproduction System for Professional Popular Music Production", 28th VDT International Convention, Cologne, Germany, November 2014.

[7] Lindau A.; „Binaural Resynthesis of Acoustical Environments“, Dissertation, Technische Universität Berlin, 2014.

[8] McKee, J.: „Is it Live or is it Edison?“, Library of Congress, https://blogs.loc.gov/now-see-hear/2015/05/is-it-live-or-is-it-edison/, Stand 10.7.2017.

[9] Rodigast R., Seideneck M., Frutos-Bonilla J., Gehlhaar T., Sladeczek C.; „Objektbasier- te Interaktive 3D Audioanwendungen“, Fachzeitschrift für Fernsehen, Film und Elekt- ronische Medien FKT, Ausgabe 11, November 2016.

[10] Scuda U., Stenzel H., Baxter D.; „Using Audio Objects and Spatial Audio in Sports Broadcasting“, 57th AES Conference: On the Future of Audio Entertainment Technology, Hollywood, USA, March 2015.

[11] Snow, W.B.; „Basic Principles of Stereophonic Sound“,IRE Transcaction on Audio, 3(2):42-53.

[12] Verheijen, E.; „Sound Reproduction by Wave Field Synthesis“, Dissertation, Delft University of Technology. 1997.

Digital Radio

6

Worldwide premier radio quality

Prof. Dr. Albert Heuberger
Fraunhofer Institute for Integrated Circuits IIS

Summary

Digitization is moving ahead at full speed. For most people, the smartphone is a constant companion; manufacturing businesses are also pushing ahead with digitization on the factory floor in the wake of Industry 4.0. Even radio cannot stop the trend in its tracks: step by step, digital radio is replacing its familiar FM cousin. A common practice in numerous European nations, and many developing nations are preparing for the switchover. Digital radio offers numerous advantages: greater program diversity, improved reception, a wealth of enhanced services. Digital radio will grow together with mobile telephony in the long term. Whereas radio sends information that is of interest to everyone, mobile telephony takes on “personalized” information. In this way, the two technologies can be the ideal complement to one another.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_6

6.1 Introduction

Crackling and hissing on the radio? No way! Digital radio consigns that sort of interference to history. It offers listeners and radio stations alike numerous advantages. Listeners gain more program diversity along with supplementary information via data services. Reception quality, too, is better. Radio stations in turn save energy due to increased transmission efficiency and are at the same time able to broadcast a greater number of programs due to the more efficient use of the transmission spectrum. Both provide economic benefits. Just like familiar FM radio, terrestrial digital radio uses radio signals; but contrary to popular belief, there is absolutely no need for an Internet connection, making it thus free to receive.

In most European nations, DAB+ digital radio is already part of everyday life. Norway is relying exclusively on this new form of radio – it switched FM radio off by the end of 2017. Switzerland and the UK, too, are actively considering moving away from FM before 2020. Numerous developing nations are currently planning the transition from analog short- and medium-wave radio to DRM digital radio, and the digitization of local FM radio infrastructure has started. India is one of the pioneers here and is on the way to becoming the largest market for digital radio in the world. All of the necessary technologies – from the necessary basic technologies to transmission and reception systems for digital radio applications – have been co-developed with significant contributions by the Fraunhofer Institute for Integrated Circuits IIS. Software built by Fraunhofer, for example, is contained in every single DRM device in India.

6.2 Spectrum efficiency allows more broadcasts

One of the great advantages of digital radio lies in its spectrum efficiency. In order to explain what lies behind this, we will first take a look at traditional FM radio. Traditional FM radio operates at frequencies of 87.5 to 108 megahertz, giving it a bandwidth of about 20 megahertz. This means that the available frequencies are very limited. A lot of radio stations that would love to enter the market are shut out because there are no more free frequencies available.

Digital radio comes to the rescue. The frequency resource is limited here, too, but in the 20-megahertz useable spectrum there is room for up to four times as many broadcasters. The exact number depends on the audio quality desired and the robustness of transmission – if a program is broadcast at a higher quality, it requires a higher bit rate. The increased efficiency is first and foremost the result of compression, that is, of the standardized xHE-AAC and HE-AAC audio codecs, developed essentially by researchers at Fraunhofer IIS. These codecs are responsible for reproducing speech and audio at lower data rates with higher quality and thus form the foundation for the high sound quality of digital radio. The second transmission quality parameter is the robustness of the signals transmitted. Depending on how much supplementary error-correction information is available for the transmission, reception of the signal is better or worse, with a direct impact, for example, on the broadcaster's range.

One additional important reason for the greater broadcast capacity is what is known as the single frequency network. With digital radio, all stations broadcasting the same information can do so using the same frequency; at the receiving end, where the signals from two transmitters are received simultaneously, there is no interference. With FM radio, this is not the case. Here, neighboring transmitters have to work on different frequencies. Imagine taking a car journey through Germany: the frequencies for a given station change as you move from one area to another. A single program thus simultaneously requires several frequencies. With digital radio, a single frequency suffices to cover the entire country. In this way, significantly more radio programs can be accommodated in a given frequency spectrum than before.

6.3 Program diversity

With FM radio, each program needs its own frequency. This means that if Bayern 3 is broadcasting on a certain frequency with a bandwidth of 200 kHz, this frequency range is occupied and is only available to other broadcasters if they are located far away. Digital radio on the other hand is more diverse – with DAB+, 15 different programs can typically be broadcast over a single frequency with a bandwidth of 1.536 MHz. So you could broadcast Bayern 1, Bayern 3, and various regional programs within the entire transmission area on the same frequency. These are known as a *multiplex* or *ensemble*. It opens up greater program diversity to listeners. Black Forest Radio ("Schwarzwaldradio"), for example, was only available in the region of the same name via FM. On the nationwide digital radio multiplex it can now be heard across the nation. People from Hamburg, for example, who like to holiday in the Black Forest can now listen to "holiday radio" at home, too. Classic Radio ("Klassikradio"), too, already has a place in the German nationwide multiplex – the broadcaster is increasingly emphasizing digital radio and has already switched off its first FM transmissions. These examples show how digital radio opens up the possibility of offering special-interest radio to listeners.

At the moment, digital radio use in Germany is around 12% to 18%; advertising revenues are thus still limited. Nevertheless, numerous broadcasters are already turning to digital radio, especially in order to secure a transmission slot for themselves for the future.

6.4 Innovative services: from traffic alerts to emergency management

Digital radio not only transmits radio content in high quality, it also offers a diverse range of new services. One example is traffic information services. Current navigation systems receive analog radio traffic updates via the traffic message channel (TMC) – a narrow-band extension to the FM signal – which are then factored into the navigation. The trouble is this: the amount of information that can be sent via TMC is severely limited. This creates difficulties when it comes to locating traffic jams, for example. In order to keep the content as brief as possible, instead of sending specific location information via TMC, so-called *location codes* are sent, which are then translated back by the receiver device from a cached list. Newly-built freeway exit-ramps, however, are not included in this list and thus cannot be used as locations. Any number of new location codes also cannot simply be added to the list. Although significantly more accurate data is available on the sending side, the limited data rate prevents it from being communicated. In short, TMC is a limited and generally overloaded channel that is no longer a fit for the demands of modern navigation devices. Higher quality navigation systems receive the information via mobile telephony, but this does not always function seamlessly either. For one, transmitting information this way is expensive, and each car also needs to be notified individually. If numerous drivers request information simultaneously, the network quickly becomes overloaded.

Digital radio enables these kinds of traffic information services to operate significantly more efficiently. The associated technology is known as TPEG, short for Transport Protocol Experts Group. Whereas the data rate for TMC is 1 kbit/s, for TPEG it is usually 32 to 64 kbit/s. Even higher data rates would not be a problem where required. In addition, the information arrives intact even under poor reception conditions. This opens up new applications such as getting an alert about the tail end of a traffic jam. Using sensor data from the freeways or via floating data from driver smartphones, precise calculations can be made regarding where a traffic jam ends. A driver alert must be sent during the right time period: if the alert arrives too early, the driver will already have forgotten about it by the time they reach the traffic jam end tail; if it arrives too late, the driver has no time to respond. The alert should thus be issued 800m to 1500m before the tail end of the traffic jam. In other words, transmission needs to be exceptionally timely and reliable. In addition, the tail end of the traffic jam also moves, so the data needs to be updated every minute. All of this is possible using TPEG and digital radio – and it is completely irrelevant how many people require this information simultaneously.

Digital radio can also be used to send traffic predictions to navigation systems. If you are setting out from Munich to Hamburg, for example, then it is of little concern to you at that point whether there is a traffic jam in Kassel. What you want to know is what will the traffic situation be like there in three to four hours? If the navigation system receives traffic predictions in Munich already, for the suggested route, it can largely bypass congestion in advance. Parking availability information would be another conceivable possibility: using digital radio, the navigation system can announce where the most parking is available. TPEG offers all of these possibilities; with regard to traffic information, it is an enormously more powerful tool than the old TMC or mobile telephony.

New digital radio services are not at all limited only to traffic. Digital radio also offers numerous applications beyond this. The news service *Journaline* for digital radio (again developed in partnership with Fraunhofer IIS researchers) allows listeners to receive comprehensive information that they can read on the radio's display. This may be background information regarding the radio program or the music being played, but it could also be airport departure times or soccer results. Journaline is like a kind of video text for radio – adapted, of course, to modern conditions. It not only features intuitive operation but also additional functionalities such as Internet links, images, and local information.

Digital radio also has significant potential in the field of public safety notices. The Emergency Warning Functionality (EWF) – that Fraunhofer IIS played a key joint role in developing – is the first to allow real-time public alerts. This is not possible via FM since it always entails delays – information arrives first on the presenter's desk here and must then be passed on by them. This can sometimes take as long as 45 minutes. And some programs are even automated – alerts here are ignored completely. With the new EWF technology, on the other hand, the emergency services control center issues the alert directly and almost immediately. Even devices that run on standby, such as alarm clocks, can be used: they automatically turn themselves on when an alert is issued. In these cases, the alerts are brief and to the point, e.g. *Keep your doors and windows closed*. With services such as Journaline, additional information can be accessed simultaneously and in parallel via radio display – with no need for an Internet connection, which may in any case no longer be available in a disaster situation. In addition, the information is provided in various languages. This way, the alert service is also suitable for those with difficulty hearing and non-native speakers.

6.5 Non-discriminatory access

One additional advantage of digital radio lies in the non-discriminatory access to information. There are two aspects to this: while content transmissions via mobile telephony are only free if you have a contract (and therefore does not offer non-discriminatory access), digital radio can be received for free. In addition, digital radio allows the dissemination of textual and visual information too, for hearing-impaired people for example.

6.6 Hybrid applications

Of course, the choice bet+ween mobile telephony and digital radio is not either/or. Rather, the two technologies will continue growing together in the long term, acting as helpful ancillaries to one another. While digital radio takes care of information that is of interest to everyone, mobile telephony looks after "personalized" information. The kind of form this could take is indicated by the Journaline service mentioned above. Current news, for example, can be transmitted via digital radio. The user can then access additional information about events in the immediate vicinity through Internet links as required; this is achieved via mobile telephony. The user is unaware of the process – everything appears as a single homogenous service.

The advantages of mobile telephony and digital radio can be combined for the ideal outcome. Mobile telephony is strong when it comes to information that interests individual users in particular. When it comes to listening to the radio, however, mobile telephony is a poor option. In the end, cellular sites only have limited total capacity. If a user is listening to the radio and moves from cell A to cell B, then cell B must immediately provide the necessary capacity. Mobile service providers need to maintain spare capacity "just in case". Mobile telephony is thus only of limited suitability for "one-to-many" applications – this kind of information belongs on broadcast-type networks that are precisely tailored to these tasks.

6.7 Outlook

In Germany new regional and local multiplexes will start in various regions.

All over Europe, there is a growing number of DAB+ programs with additional services such as the news service Journaline and slideshow images. In its Digital Radio Report, the European Broadcasting Union (EBU) finds that more than 1,200 radio stations are currently broadcasting in Europe via DAB or DAB+, of which more than 350 are only available digitally. Digital radio thus offers increasing usefulness and added value to listeners, on top of the advantages already listed such as a wider variety of programs, improved sound quality, more stable networks, as well as cost and energy savings due to more efficient frequency use and lower transmission power. Especially in countries with poor or no Internet connection, these new systems facilitate free and widespread access to information, entertainment and education.

Contact details:
Prof. Dr. Albert Heuberger
+49 9131 776-1000
albert.heuberger@iis.fraunhofer.de

5G Data Transfer at Maximum Speed

7

More data, more speed, more security

D. Eng. Ronald Freund · D. Eng. Thomas Haustein ·
D. Eng. Martin Kasparick · D. Eng. Kim Mahler ·
D. Eng. Julius Schulz-Zander · D. Eng. Lars Thiele ·
Prof. D. Eng. Thomas Wiegand · D. Eng. Richard Weiler
Fraunhofer Institute for Telecommunications,
Heinrich-Hertz-Institute, HHI

Summary

Mobile communications have permanently changed our society and our ways of communicating ever since the global availability of mobile speech services, and because these communications form the basis for mobile Internet. They have facilitated a new dimension of productivity growth and manufacturing and service process networking since the use of the Internet. Their technical basis is founded on a deep understanding of the relationships between radio and telecommunications technology, beginning with radio wave propagation and modeling, through techniques for digital signal processing and a scalable system design for a cellular radio system with mobility support, to methods for system analysis and optimization. The Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute, HHI has been working in the field of mobile telephony communications for 20 years and has made key contributions to the third, fourth, and fifth generations. Alongside research articles and numerous first-time demonstrations of key technological components, the institute is also an active contributor to 3GPP standardization.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_7

7.1 Introduction: the generations of mobile communications from 2G to 5G

The furious onward march of digitization in society requires the support of flexible and scalable mobile telephony solutions in order to meet the demands of new applications and business models. The socio-economic transformation expected is built upon the availability of mobile connections to the data network everywhere, at any time, and with consistent quality of service for communications between people, between people and machines, and between machines. This requires a fundamentally new vision of mobile communications. While the first generation of mobile telephony was founded on analog radio technology, 2G (GSM) was a purely digital communications system right from the start, which made it possible to provide speech and the first data services (SMS) globally. UMTS, the third generation (3G), enabled the technological basis for the mobile Internet via broadband mobile telephony connections. 4G extended the dimension of broadband data communications significantly and reduced its complexity by means of a so-called all-IP approach handling both speech as well as IP-based data services. At the same time, this laid the foundation for a convergence between fixed-line and mobile telephony communications networks.

The goal for the fifth generation is to enable mobile data communications in new fields as a platform for communications between everything and everyone, opening up brand new possibilities for production, transport and healthcare, and addressing issues such as sustainability, security and well-being that are relevant to modern society. The vision of a completely mobile and connected society requires the supporting conditions for an enormous growth in connectivity and traffic volume density in order to sustain the broadest possible range of use cases and business models. The long-discussed concept of the Internet of Things will finally be made possible via the organic embedding of mobile communications capabilities in every field of society. A comprehensive motivation together with the goals of 5G were first outlined in the NGMN White Paper [1]; a current analysis of the challenges, trends and initial field trials was set out in the IEEE Journal on Selected Areas in Communications [18]. Driven on by technological developments and socio-economic transformations, the 5G business context is characterized by changes in customer, technological and operator contexts [2].

Consumer perspective: The significance of smartphones and tablets will continue to grow as it has since their introduction. It is expected that smartphones will remain the most important personal devices in future, continuing their development in terms of performance and capabilities, and that the number of personal devices will increase significantly through new devices such as wearables and sensors. Assisted by cloud technology, the capabilities of personal devices will be seamlessly extended to all sorts of applications such as high-quality (video) content production and sharing, payments, proof of identity, cloud gaming, mobile television and supporting intelligent living in general. These devices will play a significant role in the fields of healthcare, security and social life, as well as in the operation and monitoring of household appliances, cars, and other machines.

The mobile telephony industry is expecting the first 5G devices with a limited range of functions in 2018 and 2020 during the Olympic Games in South Korea and Japan, with a large-scale rollout starting from 2022.

Business context: Analogous trends to those in the consumer realm will also feed into daily company life; the boundaries will blur the line, for example, between private and professional use of devices. Businesses thus need flexible and scalable solutions in order to manage the security and data protection issues that arise in this usage context. For businesses, mobile communications solutions are some of the key drivers of increased productivity. Over the coming decades, businesses will increasingly make their own applications available on mobile devices. The spread of cloud-based services facilitates the portability of applications across several devices and domains and offers entirely new opportunities, together with new challenges as regards to security, privacy, and performance.

Business partnerships: In many markets we see the trend of network operators entering into partnerships with so-called over-the-top (OTT) players, in order to provide better integrated services to their respective end customers. For OTT players the communication network's quality of service profile is becoming increasingly important; it is necessary in order to be able to provide new services in the private and above all business spheres. Inherent synergy here between connectivity with a guaranteed quality of service, on the one hand, and high-quality services on the other, enables these partnerships to become the foundation for shared success.

Alongside classical broadband access, 5G is also relevant for new markets, particularly in the area of vertical industries. The combination of sales of more than 300 million smartphones per quarter, with a total of more than 10 billion smartphones in existence globally, combined with an expected future 50–100 billion radio-connected devices (machine-to-machine communications), means that an increase in overall data traffic by a factor of 1,000 by the year 2020 can be expected. Alongside pure peak data rates, the field will be confronted with additional demands such as, for example, extreme ranges, extreme energy efficiency, ultra-short latencies, extremely high availability or reliability, through to scaling issues with massive user access, massive antenna systems, and heterogeneous radio access and convergence requirements for the networks.

The breadth of application scenarios to be serviced demands a variety of communications solutions in radio access and at the network layer, which, despite their heterogeneity, need to be incorporated into the 5G framework via standardized interfaces.

This highly dynamic environment provides Fraunhofer with opportunities to actively contribute to the innovation process, to develop new technological solutions and thus make a significant contribution to the ecosystem. Due to the shift in the focus of applications away from people as the end customer (in the case of telephony or the mobile Internet), business-to-business (B2B) communications solutions will in future increasingly become the focus as the enabler of automated processes. The demands are varied and cannot be fulfilled with a one-size-fits-all solution. New market participants from the Internet sphere are flooding into the field and significant transformation of the existing mobile telephony ecosystem is to be expected. Fraunhofer's understanding of sector-specific challenges provides it with opportunities here to develop targeted solution components for and with clients and partners.

From pure theory through to initial field trials (proofs-of-concept), the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute, HHI has already made significant contributions to 4G-LTE research and the continued development of LTE Advanced and LTE Advanced Pro, as well as in the early phases of 5G research. These have included extensive studies on *wave forms* [32][38][42][41][28], *MIMO* [37][24][19], *CoMP* [34][35][31][36], *Relaying* [33][40], *cognitive spectrum management* [CoMoRa][29][30], *energy-efficient network management* [43] and other key technologies [39][27][20] that today form the starting point for new approaches to 5G.

7.2 5G vision and new technological challenges

Within 5G research, it is particularly important to address the areas of application that have hitherto only been able to benefit in limited manner from the existing opportunities provided by mobile telephony, and which may in the future represent particular growth markets for 5G.

5G thus becomes the door to new opportunities and use cases, of which many are as yet unknown. Existing mobile telephony standards permit smartphone connectivity that will reach even higher data rates via 5G. Alongside connecting people, 5G will additionally facilitate the connection of intelligent objects such as cars, household appliances, clocks, and industrial robots. Here, many use cases will make specific demands on the communications network with respect to data rates, reliability, energy consumption, latency etc. This diversity of applications and the corresponding requirements will necessitate a scalable and flexible communications network, as well as the integration of diverse and partly very heterogeneous communications solutions.

Vertical markets: The fifth wave of mobile communications is intended to make industry and industrial processes more mobile, and to automate them. This is often referred to as machine type communication (MTC) or the Internet of Things (IoT). Between 10 billion and 100 billion intelligent devices with embedded communications capabilities and integrated sensors will be enabled to respond to their local environment, communicate across distances, and interact in control loops as the basis for complex multi-sensor multi-actor systems that could previously only be realized as wired systems. These devices have a heterogeneous spectrum of requirements in terms of performance, power consumption, lifetime, and cost. The Internet of Things has a fragmented spectrum of communications requirements with regard to reliability, security, latency, throughput, etc. for various applications. Creating new services for vertical industries (e.g. health, automobiles, home, energy) often requires more than pure connectivity, but also requires combinations with, for example, cloud computing, big data, security, logistics and/or additional network capabilities.

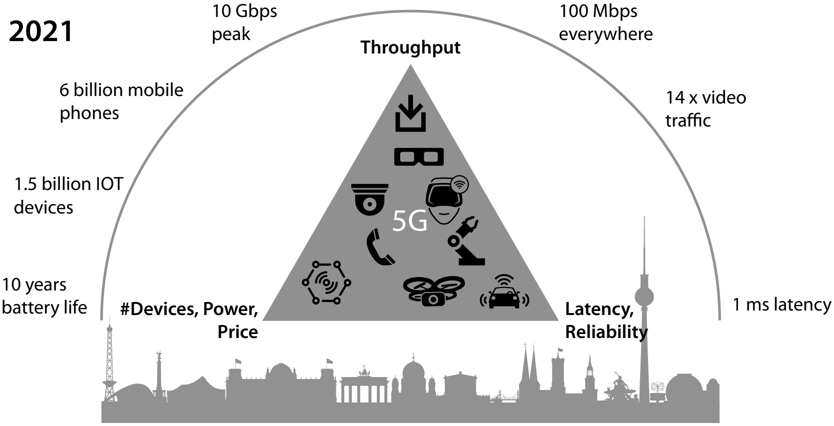


Fig 7.1 Classification of the most important areas of application for 5G (image source: ITU-R IMT2020, Fraunhofer HHI)

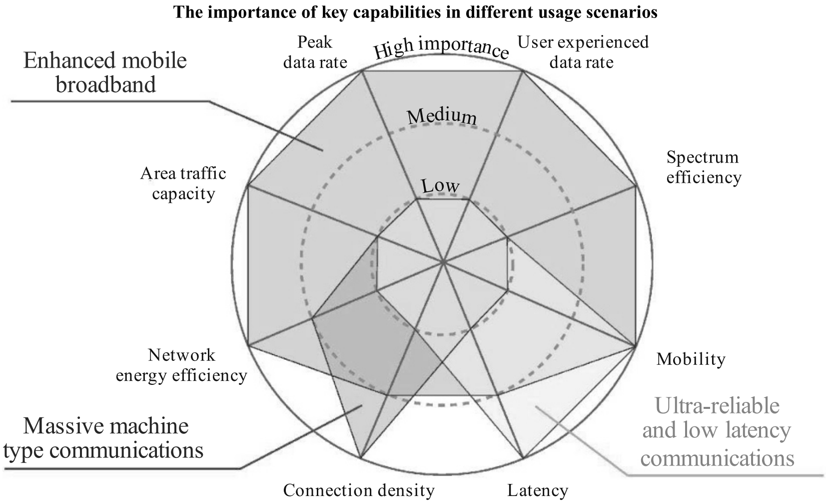


Fig 7.2 KPIs for the application categories; markings at the outer edge indicate high importance, further inside lesser importance (ITU-R IMT2020, Fraunhofer HHI)

The diversity of the areas of application and scenarios for 5G that were discussed as relevant can be categorized into three main groupings from a communications requirements point of view, as shown in Fig. 7.1. The relevant key performance indicators (KPIs) indicating the importance of each of the three categories are illustrated in Fig 7.2.

Broadband mobile telephony scenarios with the highest data rates

Enhanced Mobile Broadband (eMBB) has the goal to increase the data rate one thousand-fold per area or a data rate of 100 Mbps available everywhere. This will enable high-resolution video data and applications such as virtual or augmented reality to even become mobile possibilities for the first time.

The thousand-fold increase of the available communications data rate per area requires new approaches in the design of the communications infrastructure and of the transfer mechanisms used. Aspects such as an improved energy efficiency per bit or the reduction in system self-interference must be overcome through new solution strategies.

Radio solutions with low latency and high reliability

Mission-critical control mechanisms require reliable, low-latency communications. This is necessary in order to facilitate security-critical applications such as in Industry 4.0 scenarios or in the case of automated driving. For the latter, for example, a highly reliable radio connection is needed to avoid a critical system state developing due to “dead spots” or lost packets, potentially resulting in an accident being caused. Also, low delay times are necessary so that moving objects such as vehicles are able to quickly react to dangerous situations. These kinds of applications are often referred to as *tactile Internet* [22][26] and generally comprise control loops with low latencies and radio-based communication paths.

Up to now, communications solutions were designed for wide area networks, and latencies tailored to the needs of the users (people) at around 10 ms to 100 ms.

Machine networking enables us to operate complex feedback control mechanisms via radio connections that must fulfill latency requirements for the control loop time constant of 1 ms or less. This necessitates a completely new design for many components in the communications pathway [25][23] and a further tendency towards distributed and local signal processing.

Massive connectivity – the Internet of Things

The task of Massive Machine Type Communications (mMTC) is to connect billions of objects in order to create the Internet of Things and facilitate diverse new kinds of applications. Sensor data can provide a range of benefits in smart cities, for example, through the use of narrow-band, energy-saving communications.

Existing communications access networks were and are designed for today’s typical user numbers (people and computers) per area. Connecting sensor networks and developing an Internet of Things requires both completely new scaling options in the number of end devices per radio cell/area as well as simultaneous access to the shared medium of the mobile telephony spectrum.

Due to the diversity of the new possibilities that are expected via 5G, this chapter will only go into detail on a select number of areas of application, in particular automated driving and Industry 4.0.

Intelligent traffic and logistics

New kinds of traffic and logistics solutions are one example of an important 5G application that we will take a look at here. Classical mobile communications take place between end devices (such as smartphones) and a mobile telephony infrastructure. Where there are high vehicle densities and speeds, for example, together with a requirement for local communications between vehicles in close proximity, neither classical mobile telephony solutions nor WLAN systems are suitable for facilitating reliable inter-vehicle communications. New kinds of approaches to ad hoc mesh networking with the required scalability need to be developed and integrated into the existing radio systems. The high mobility of the radio subscribers and the resulting radio resource allocation dynamic necessitate entirely new systems approaches with respect to scalability, cognition, and resilience in their self-configuration and optimization.

Although radio communication is not a necessary precondition for automated driving, it will nevertheless play a key role in particularly complex environments. Sharing on-board sensor data with neighboring vehicles can dramatically extend a vehicle's perception and open up the possibility of automated driving in urban environments for the first time. This application of 5G requires high data rates of up to 1 Gbps (eMBB) and simultaneously high reliability/low latency (URLLC).

The diversity of 5G applications means that we expect different players from previously unrelated industries to coordinate their communications requirements across sectors. The 5GAA (5G Automotive Association) [44] is an example of an organization formed from the automotive and telecommunications industries. Due to the technological convergence of mobile communications devices and communicative automobiles that we can expect, the founding of 5GAA is a logical step, and it means that we can expect fruitful synergies between these two sectors in future.

5G thus considers itself as an end-to-end ecosystem to facilitate a completely mobile and connected society. 5G signifies the integration of different networks for various sectors, domains, and applications. It enables the creation of added value for customers and partners via new, sustainable business models.

7.3 Technical key concepts: spectrum, technology and architecture

Every generation of mobile telephony requires dedicated key technologies in order to meet the necessary requirements in terms of data rate, reliability, energy efficiency, etc. Examples here are multi-antenna and millimeter wave technologies, interference management, cognitive and flexible spectrum access, and the corresponding management. Fraunhofer HHI has been working for 20 years with partners from industry and research on suitable mobile telephony solutions and is active in the development, testing, and standardization of new 5G solution components.

Multi-antenna technologies: evolution from 4G: massive MIMO

Ever since 3GPP LTE Release 10, a large number of antennas are applied in fourth generation mobile radio systems. These technologies are known in specialist circles

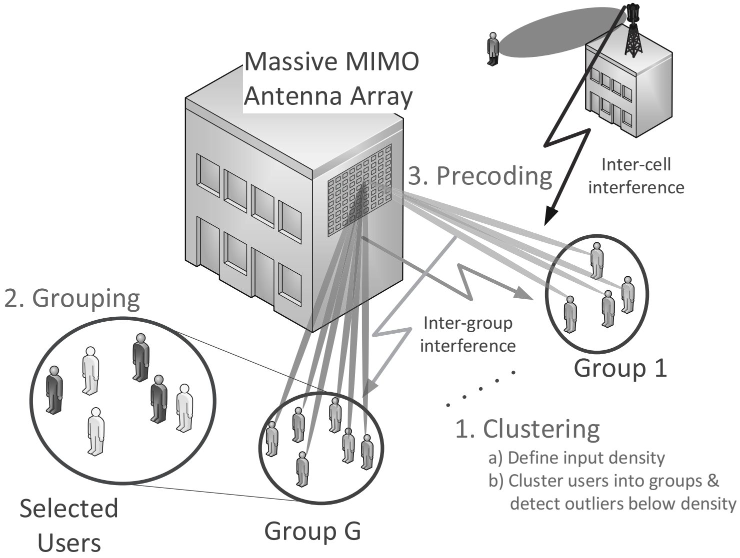


Fig. 7.3 Principle of the multilevel precoding process (Fraunhofer HHI)

as *full dimension MIMO* and are expected to resolve the need for ever-increasing peak data rates in radio cells. This is facilitated via multiuser multiplexing mechanisms – that is, supporting a range of users over the same time and frequency resources by utilization of the spatial dimension.

Fig. 7.4 Examples of a 64-element planar (left) and 128-element circular antenna array (right) for massive MIMO antennas for new radio bands at 3.5 GHz (Fraunhofer HHI)

In the fifth generation mobile telephony standard, the number of active antenna elements is expected to once again be significantly increased; this is known as *massive MIMO*, where data rates are significantly above 1 GBit/s. This necessitates major changes with respect to the current standard in order to guarantee cost-efficient operation. Multilevel spatial beamforming methods need to be extended so that the large number of antenna elements can be divided into so-called sub-antenna arrays. In each of these sub-antenna arrays, beamforming weights are then adapted in keeping with the phase, whereas phase control of different sub-antenna arrays must only be readjusted or coordinated on a slow time base [6]. New kinds of antenna shapes play an important role in meeting the heterogeneous requirements in the cellular system. Thus so-called planar antenna arrays can be utilized, for example, in order to shape a range of differing beamformers in a clearly defined solid angle [3]. In contrast, (semi-)circular antenna arrays may ideally be used in order to

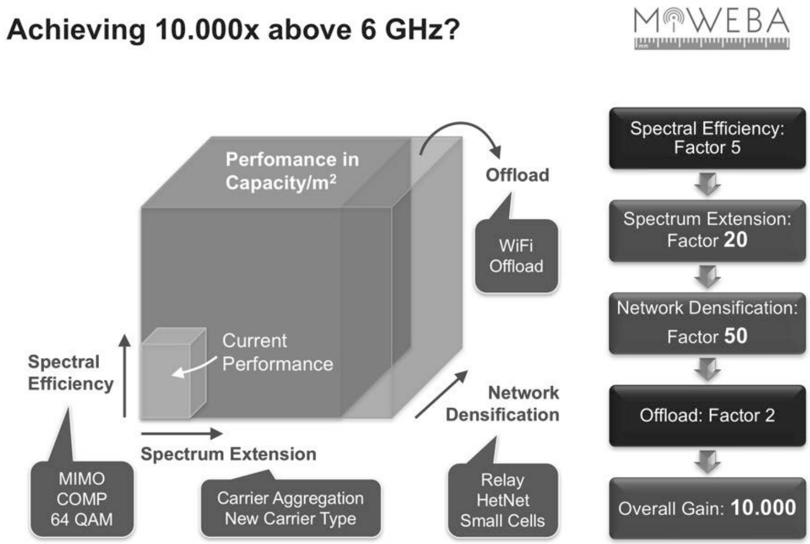


Fig. 7.5 Illustration of the degrees of freedom for realizing an increase in capacity by a factor of 10,000 compared to existing 4G solutions. A combination of network densification (ultra-dense networks) and the use of a spectral region above 6 GHz demonstrates the necessary potential (source: MiWEBA, joint EU/Japanese research program as part of FP7). Within the framework of 5G standardization, a distinction is currently being made between solutions below and above 6 GHz; within the 6–100 GHz band, the regions around 28 GHz, 60 GHz, and 70 GHz are initially being addressed for reasons of spectral region availability and technological maturity [21]. (Fraunhofer HHI)

evenly illuminate a wide angular range and to simultaneously achieve a variable sectorization [4].

Massive MIMO allows the precise spatial discriminability of radio signals in the angular range so that the positions of end devices (which are no longer necessarily represented by people in the 5G standard) can be estimated at high-precision in order to ideally only send the data to the end device in question and not unnecessarily cover the environment with disturbance power. Gathering precise positional data for all autonomous aircraft and vehicles is vital, for example. In typical cellular systems, positioning via mobile telephony can be improved from today's accuracy of approximately 50 m to under 1 m, without using so-called GNS systems (Global Navigation Satellite System). One new application of this potential being researched by the MIDRAS project [5] is the spatial detection and targeted jamming of unauthorized civilian micro drones via distributed massive MIMO antenna systems.

5G New Radio: millimeter wave communication

The available spectrum is a limited resource and thus sets limits on the scalability of existing mobile telephony systems with respect to bandwidth and provided data rates. Facilitating a one-thousand-fold increase in capacities per area not only requires an increase in spectral efficiency and greater reuse of the same spectrum at different locations (spectral reuse) by introducing small radio cells; it inevitably also

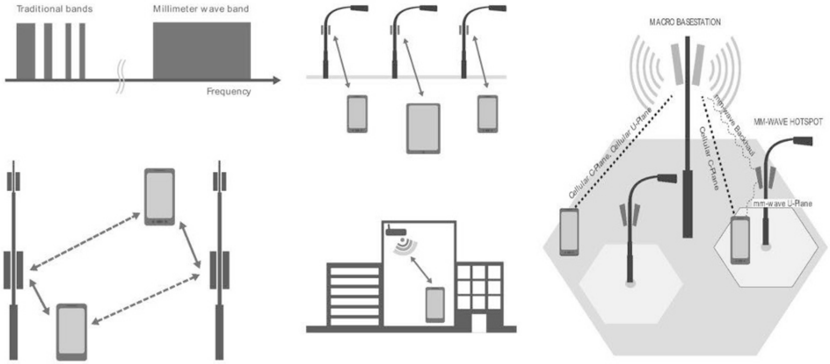


Fig. 7.6 5G key solutions for radio communications in the millimeter wave spectrum; top left: use of the spectrum above 6 GHz; bottom left: interference management between base stations and end devices within the radio ranges; top middle: densification of networks with small cells; bottom middle: small cells in indoor areas; right: incorporation of millimeter wave small cells into the macrocellular infrastructure – control plane/user plane splitting (Fraunhofer HHI)

entails using more of the spectrum. Here, the frequency range above 6 GHz to 100 GHz in particular was identified in order to enable an extension of the available spectrum by a factor of 20 [7][8].

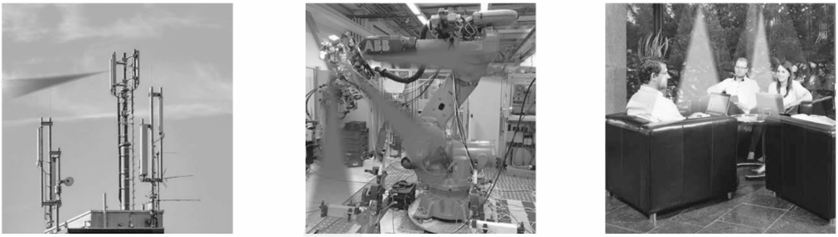
The use of spectral regions above 6 GHz requires new technologies and solutions for transceiver chips, antenna design and signal processing in order to develop cost- and energy-efficient components suitable for the mass market. Alongside technological development challenges, a deep understanding of the propagation behavior of radio waves at high frequencies is essential for a sustainable system design. Fraunhofer is contributing to a deeper understanding of wave propagation for relevant indoor and outdoor scenarios with a range of radio field measurements and is providing channel modeling contributions within the 3GPP standardization process [9].

Communication in the millimeter wave range requires a high degree of beam focusing in the direction of the communication between the end device and the base station due to the high path loss and associated restrictions in the radio range. New forms of compact antenna arrays using hybrid beamforming approaches are expected to facilitate the necessary gains here in terms of range, signal quality and interference limitation. The high frequencies permit compact integration in a limited space, but also require new approaches to connection establishment, link optimization, and link tracking, particularly in mobile scenarios.

Optical 5G wireless communications

At the moment, the current 5G discussion is focused on carrier frequencies in the lower gigahertz region, in particular for areal provision of fast network access for mobile end devices in cities and rural regions. Frequency regions up to 100 GHz are also being developed for multi-Gbps transfers, with the aid of millimeter wave technology, which is a key element for data transfer in small radio cells and for network densification.

A logical extension for 5G is the use of carrier frequencies in the terahertz region, where the electromagnetic waves propagate visibly in the form of light, or in the infrared wavelength band. To extend the 5G infrastructure, the use of LED lighting elements both indoors (ceiling lights, standing lamps, etc.) as well as outdoors (vehicle headlights, streetlights, traffic lights, etc.) for information transfer and navigation [Grobe] is very attractive. Communication via light can be considered as secure since the information can only be received within very limited lightspots. The diameter of these lightspots can be varied in size by choosing suitable optics (affordable plastic lenses), which facilitates adaptation to different application scenarios. By implementing appropriate handover mechanisms between several optical spots as well as to neighboring radio cells, a mobile communication across the area



Mobile Backhaul

Industrial Wireless

Public/Private Premises

Fig. 7.7 Application scenarios for LED-based wireless communications (Fraunhofer HHI)

can be achieved. In addition, light is highly resilient to electromagnetic interference radiation and can be used with the current 5G frequency regions without interference.

Using conventional LEDs (installed for lighting) for optical wireless communications has already been demonstrated by Fraunhofer HHI, for data rates of more than 1 Gbit/s in bidirectional application scenarios. To achieve this, real-value OFDM is used at the air interface for the transfer process [11]. Current chipsets allow the transfer of 2.5 Gbit/s per color within a bandwidth of 200 MHz, where the data rate is dynamically adapted to the quality of the transfer channel and thus can also support non-line-of-sight scenarios, in other words only via the use of reflected light. With modulation bandwidths of up to 300 MHz, LEDs are great value components for the aforementioned 5G application scenarios (see Fig. 7.7), e.g. for networking robots in industrial production, for equipping conference and school rooms with fast optical WLAN, or for 5G backhaul systems with ranges up to 200 m [Schulz].

Network slicing and convergence

Alongside innovations for the physical interfaces, 5G will also have an effect on network operation and the management of the physical resources. The use cases described feature significant differences in the peculiarities of their end-to-end requirements. Network slicing is a concept where the physical resources are abstracted and arranged as logical resources according to demand. In the access field there will be a merging of previously separate infrastructures for fixed-line and mobile network access towards a software-driven universal hardware infrastructure where various applications (e.g. autonomous driving, Industry 4.0, telemedicine, etc.) will be able to be configured and operated with their respective user groups. Network providers can thus create a slice for each specific application tailored to the respec-

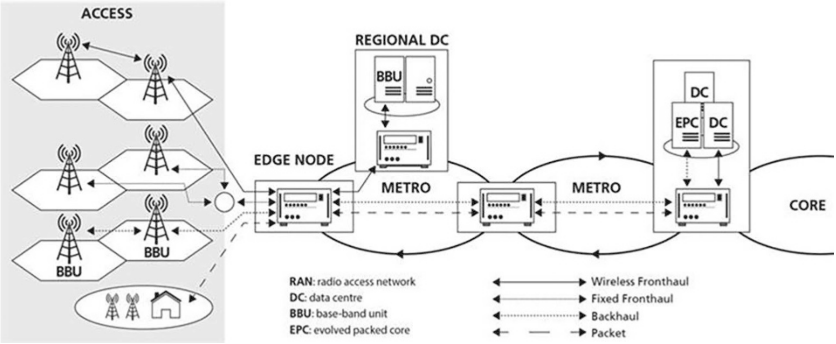


Fig. 7.8 The 5G infrastructure investigated as part of the 5G-PPP Metro-Haul project (Fraunhofer HHI)

tive requirements, producing a network-as-a-service. A slice with extremely high reliability and guaranteed latency can thus be formed for networking automated vehicles, for example, and a slice with sufficiently high data rates and limited requirements in terms of latency, resilience and availability can be provided for mobile video streaming.

The primary criteria for parametrizing the network slices are geographical coverage, connection duration, capacity, speed, latency, resilience, security, and the required availability of the communication. In order to implement network slicing, techniques from software defined networking (SDN), network function virtualization (NFV), and network orchestration are used.

From an end-to-end perspective, 5G is a fiber optics fixed network with high bit rate mobile interfaces [13]. The low-latency requirements of < 1 ms for the 5G application scenarios discussed require entirely new access and metro networks (see Fig. 7.8). Due to the variety of radio cells (including smaller ones) supported by 5G, powerful back- and front-haul system technologies are required that also facilitate a dramatic reduction in capex and opex during the construction and operation of 5G network infrastructure.

The implementation of 5G key performance indicators also has significant implications for the optical metro network, since (i) higher capacities must be transferred via this same fiber optics infrastructure, and (ii) a latency-aware metro network is required where, for example, latency-sensitive slices at the edge of the metro network (edge node) can be handled in such a way that end-to-end latency can be guaranteed. The transmission technology for broadband linking of data centers is being extended from 100 Gbit/s to 400 Gbit/s per wavelength. Additionally, a SDN-based network management will also facilitate the fast configuration of 5G services: current goals are 1 min. for simple network path establishment, 10 min. for complete installation of a new virtual network function (VNF), and 1 hour for establishing a new virtual network slice. Fig. 7.8 gives an overview of the network infrastructure for providing future 5G services addressed in research projects.

7.4 5G research at Fraunhofer HHI

The broad spectrum of open research questions that remain to be answered in the 5G context has led to a variety of research programs being initiated in Germany, Europe, and across the world in order to address the issues in good time, in a focused manner, and in partnership between industry and research. Fraunhofer HHI is engaged in numerous association and research projects related to 5G, first and foremost in the context of H2020 5GPP, as well as in various programs sponsored by German federal ministries, in particular the BMBF, BMWi and BMVI. The following sections are intended to provide a brief insight into the technical questions addressed by the Heinrich Hertz Institute in several selected 5G projects, working with partners on solutions which will subsequently, through the standardization process, deliver a lasting contribution to the fifth generation of mobile telephony.

Transfer Center 5G Testbed at the Berlin Center for Digital Transformation

The Transfer Center 5G Testbed was established at Fraunhofer HHI as part of the Berlin Center for Digital Transformation. At the 5G Testbed, ongoing research and development work into the further development of the fifth generation mobile telephony is being integrated into early trials with partners from the mobile telephony industry.

The Berlin Center was established by the four Berlin-based Fraunhofer institutes, FOKUS, HHI, IPK, and IZM, and in addition to the 5G Testbed comprises three further transfer centers: Hardware for Cyber Physical Systems, Internet of Things, and an Industry 4.0 Lab. In partnership with the universities of Berlin and Brandenburg, it thus forms a single location with strategically important core competencies in the area of digital transformation. Areas of application in the Berlin Center are Connected Industry and Production, Connected Mobility and the City of the Future, Connected Health and Medicine, and Connected Critical Infrastructures and Energy.

Even today, HHI and the 5G Transfer Center are a solid part of international projects within the Horizon 2020 research initiative of the European Commission and the 5G Infrastructure Public Private Partnership (5GPPP), including mmMAG-

IC, Fantastic 5G, 5G-Crosshaul, Carisma, One5G, and in the joint EU/Asia projects MiWEBA, STRAUSS, 5G-MiEdge, 5G!PAGODA and 5GCHAMPION, as well as in national research initiatives including Industrial Communication for Factories (IC4F), 5G NetMobil, AMMCOA or SEKOM.

IC4F – Industrial Communications for Factories

For many Industry 4.0 applications, high reliability and low radio transmission latencies are indispensable. The LTE (4G) and previous generations of mobile telephony standards do not meet these requirements. Accordingly, the forthcoming fifth generation mobile telephony standard is of paramount importance for the fourth industrial revolution, where the focus includes the secure, low-latency and reliable networking of machines. In contrast to its predecessors, 5G technology is being developed with a view to the requirements of vertical industries such as the automation industry; leading to high reliability and low latencies being the subjects of research, among others, in the 3GPP Ultra Reliable Low Latency Communication (URLLC) use case [26].

Already, laboratory trials [presse_1] have shown that a throughput of 10 Gbit/s with a latency of one millisecond and high reliability is possible. In order to achieve these values in real-world conditions, however, data transfer must become even more stable with respect to the kind of interference that may result from a variety of mobile devices and from fast-moving mobile devices. One example of an Industry 4.0 application that is not possible with 4G is control loops. These may be necessary to control a robot, for example. If the latency is too high or the reliability too low then it will not be possible to control the robot fast enough or safely enough. If, on the other hand, these requirements are fulfilled then safe, real-time control is possible as if someone were standing right beside the robot and operating it with a joystick.

Furthermore, future communications networks also need to guarantee the isolation of different data traffic in order to not endanger critical applications. This requires a holistic view of the different wireless network access technologies, backbone infrastructure and cloud resources. In particular, network functions need to be able to be dynamically placed in the network according to requirements. In this respect, technologies such as software defined networking (where network components are controlled via software) and network function virtualization (where network functions are implemented in software and dynamically placed) play an important role. The combination of these technologies is currently being researched in the BMWi-sponsored lighthouse project Industrial Communication for Factories (IC4F) [16]. The goal of this project is the development of a toolbox of technologies based on secure, resilient and real-time-capable communications solutions for the processing industry.

5GNetMobil – communication mechanisms for efficient vehicle communications

Low-latency and highly reliable data transmission is a key prerequisite for enabling many of the use cases and applications envisaged for 5G – in industrial factory automation, virtual presence, and autonomous driving especially.

In the 5GNetMobil research project, Fraunhofer HHI is working with others on developing an all-encompassing communications infrastructure for tactile connected driving. Tactile connected driving is expected to facilitate a range of improvements in traffic safety, traffic efficiency, and pollution compared with driving based exclusively on local sensor data.

Implementing this and other visions of the tactile Internet requires secure and robust communications for steering and control in real time. This necessitates a range of new solution approaches, both for dramatically reducing latency as well as for prioritizing mission-critical communications compared to classical broadband applications.

In particular, new and forward-looking mechanisms are required to ensure the cooperative coexistence of different mobile telephony subscribers, also through timely provision of the necessary network resources. Fraunhofer HHI is researching an implementation of this kind of "proactive" resource allocation by incorporating the broadest possible range of contextual information ("context awareness"), not least to facilitate demand forecasts as well as forecasts of the availability of radio resources. In this way, a dynamic 5G network configuration and optimization is

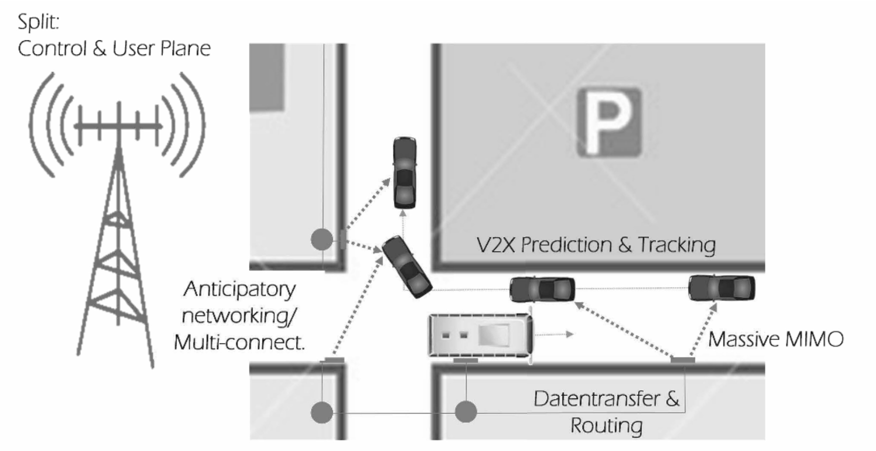


Fig. 7.9 5G mechanisms for supporting tactile connected driving (Fraunhofer HHI)

enabled, and foundations are laid for flexible decision-making. To support the high mobility requirements of tactile connected driving new, efficient and scalable cognitive network management concepts must be developed. To do this, learning algorithms are being utilized that are able to merge and process all of the available information and measurement data together with any available contextual information. One significant challenge here is guaranteeing the resilience and scalability of the outlined mechanisms even when there are a large number of users and correspondingly large quantities of sensor data and contextual information. The high reliability requirements of tactile connected driving make the use of new diversity concepts such as multi-connectivity necessary. One additional focus of studies at Fraunhofer HHI within the 5GNetMobil project is thus the development of and research into new diversity and network management strategies. In order to put network management in the position of supporting the formation of mobile virtual cells for interruption-free handovers at high mobility, new mechanisms for network coding, in-network data processing and distributed decision-making are thus being employed.

AMMCOA – 5G islands for vehicle communication beyond highways

The operation of construction and agricultural machinery is subject to especially high requirements in terms of efficiency, precision and safety. Unique features of this area of application include the unavailability of digital maps, the need to provide very high relative and absolute localization, the very high significance of coordinated use of vehicle fleets, and the need to provide a local 5G communications infrastructure (5G islands) that can function both autonomously and incorporated into wide area networks, even when there is insufficient radio network coverage from network operators. Fraunhofer HHI is working with partners in the BMBF's AMMCOA project on solutions for this application complexity. Based on its longstanding experience with millimeter wave transmission and measuring technology, HHI

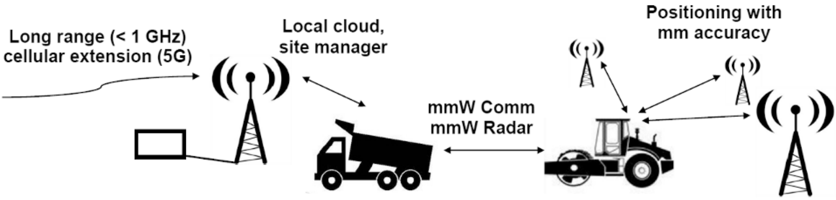


Fig. 7.10 Functionalities and basic principle of the BMBF's AMMCOA project for highly reliable and real-time-capable networking of agricultural and construction machinery, based on millimeter wave radio for communications and localization (Fraunhofer HHI)

is developing an integrated communications and localization solution with very high data rates and a localization accuracy of just a few centimeters. This solution is being integrated with additional technology components in an on-board unit for construction and agricultural machinery, and demonstrated in real-world environments by application partners from the consortium.

7.5 Outlook

In this chapter, we have provided a brief insight into the current focus areas of fifth generation mobile telephony research. Alongside the current 5G standardization process in 3GPP, not only fundamental but also implementation- and praxis-driven questions remain to be answered. The Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, HHI is working in association with other Fraunhofer institutes on technological solutions that are able to be included as key components in the overall specification for 5G, or as use case-specific or scenario related solution modules for specific industries. Due to its close cooperation with customers from the broadest range of sectors, Fraunhofer is able to work across sectors and disciplines here, making a significant contribution to the 5G ecosystem.

Sources and literature

[1] NGMN 5G White Paper, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf
[2] https://www.ericsson.com/research-blog/lte/release-14-the-start-of-5gstandardization
[3] G. Fodor et al., „An Overview of Massive MIMO Technology Components in METIS," in IEEE Communications Magazine, vol. 55, no. 6, pp. 155-161, 2017.
[4] One5G, https://5g-ppp.eu/one5g/
[5] MIDRAS: Mikro-Drohnen-Abwehr-System, https://www.hhi.fraunhofer.de/abteilungen/wn/projekte/midras.html
[6] M. Kurras, L. Thiele and G. Caire, „Interference Mitigation and Multiuser Multiplexing with Beam-Steering Antennas", WSA 2015; 19th International ITG Workshop on Smart Antennas; Proceedings of, pp. 1-5, March. 2015
[7] Millimeter-wave evolution for 5G cellular networks, K Sakaguchi, GK Tran, H Shimodaira, S Nanba, T Sakurai, K Takinami, IEICE Transactions on Communications 98 (3), 388-402
[8] Enabling 5G backhaul and access with millimeter-waves, RJ Weiler, M Peter, W Keusgen, E Calvanese-Strinati, A De Domenico, Networks and Communications (EuCNC), 2014 European Conference on, 1-5

[9] Weiler, Richard J., et al. „Quasi-deterministic millimeter-wave channel models in MiWEBA.“ EURASIP Journal on Wireless Communications and Networking 2016.1 (2016): 84.

[10] L. Grobe et al., „High-speed visible light communication systems,“ IEEE Comm. Magazine, pp. 60-66, Dec. 2013.

[11] V. Vucic et al., „513 Mbit/s Visible Light Communications Link Based on DMT-Modulation of a White LED,“ Journal of Lightwave Technology, pp. 3512-3518, December 2010.

[12] D. Schulz et al., „Robust Optical Wireless Link for the Backhaul and Fronthaul of Small Radio Cells,“ IEEE Journal of Lightwave Technology, March 2016.

[13] V. Jungnickel et al., „Software-defined Open Access for flexible and service-oriented 5G deployment,“ IEEE International Conference on Communications Workshops (ICC), pp. 360 – 366, 2016.

[14] J. Fabrega et al., „Demonstration of Adaptive SDN Orchestration: A Real-Time Congestion-Aware Services Provisioning Over OFDM-Based 400G OPS and Flexi-WDM OCS,“ Journal of Lightwave Technology, Volume 35, Issue 3, 2017.

[15] https://www.hhi.fraunhofer.de/presse-medien/nachrichten/2017/hannover-messe-2017-ultraschneller-mobilfunkstandard-5g.html

[16] IC4F Projekt: https://www.ic4f.de

[17] Cognitive Mobile Radio, BMBF Projekt 2012-2015, http://www.comora.de

[18] M. Shafi et. al. „5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment and Practice,“ in IEEE Journal on Selected Areas in Communications, April 2017, DOI:10.1109/JSAC.2017.2692307.

[19] M. Kurras, L. Thiele, T. Haustein, W. Lei, and C. Yan, „Full dimension mimo for frequency division duplex under signaling and feedback constraints,“ in Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016, pp. 1985–1989.

[20] R. Askar, T. Kaiser, B. Schubert, T. Haustein, and W. Keusgen, „Active self-interference cancellation mechanism for full-duplex wireless transceivers,“ in International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom), June 2014.

[21] K. Sakaguchi et.al., Where, When, and How mmWave is Used in 5G and Beyond, arXiv preprint arXiv:1704.08131, 2017

[22] ITU-T Technology Watch Report, „The Tactile Internet,“ Aug. 2014.

[23] J. Pilz, M. Mehlhose, T. Wirth, D. Wieruch, B. Holfeld, and T. Haustein, „A Tactile Internet Demonstration: 1ms Ultra Low Delay for Wireless Communications towards 5G,“ in IEEE INFOCOM Live/Video Demonstration, April 2016.

[24] T. Haustein, C. von Helmolt, E. Jorswieck, V. Jungnickel, and V. Pohl, „Performance of MIMO systems with channel inversion,“ in Proc. 55th IEEE Veh. Technol. Conf., vol. 1, Birmingham, AL, May 2002, pp. 35–39.

[25] T. Wirth, M. Mehlhose, J. Pilz, R. Lindstedt, D. Wieruch, B. Holfeld, and T. Haustein, „An Advanced Hardware Platform to Verify 5G Wireless Communication Concepts,“ in Proc. of IEEE VTC-Spring, May 2015.

[26] 3GPP TR 36.881, „Study on Latency Reduction Techniques for LTE,“ June 2016.

[27] R. Askar, B. Schubert, W. Keusgen, and T. Haustein, „Full-Duplex wireless transceiver in presence of I/Q mismatches: Experimentation and estimation algorithm,“ in IEEE GC 2015 Workshop on Emerging Technologies for 5G Wireless Cellular Networks – 4th International (GC’15 – Workshop – ET5G), San Diego, USA, Dec. 2015.

[28] Dommel, J., et al. 5G in space: PHY-layer design for satellite communications using non-orthogonal multi-carrier transmission[C]. Advanced Satellite Multimedia Systems Conference and the 13th Signal Processing for Space Communications Workshop (ASMS/SPSC), 2014 7th. 2014: Livorno. p. 190-196.

[29] M. D. Mueck, I. Karls, R. Arefi, T. Haustein, and W. Keusgen, „Licensed shared access for wave cellular broadband communications," in Proc. Int. Workshop Cognit. Cellular Syst. (CCS), Sep. 2014, pp. 1–5.

[30] M. Mueck et al., „ETSI Reconfigurable Radio Systems: Status and Future Directions on Software Defined Radio and Cognitive Radio Standards," IEEE Commun. Mag., vol. 48, Sept. 2010, pp. 78–86.

[31] J. Dommel, P.-P. Knust, L. Thiele, and T. Haustein, „Massive MIMO for interference management in heterogeneous networks," in Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 IEEE 8th, June 2014, pp. 289–292.

[32] T. Frank, A. Klein, and T. Haustein, „A survey on the envelope fluctuations of DFT precoded OFDMA signals," in Proc. IEEE ICC, May 2008, pp. 3495–3500.

[33] V. Venkatkumar, T. Wirth, T. Haustein, and E. Schulz, „Relaying in long term evolution: indoor full frequency reuse", in European Wireless, Aarlborg, Denmark, May 2009.

[34] V. Jungnickel, T. Wirth, M. Schellmann, T. Haustein, and W. Zirwas, „Synchronization of cooperative base stations," Proc. IEEE ISWCS '08, pp. 329 – 334, oct 2008.

[35] V. Jungnickel, L. Thiele, M. Schellmann, T. Wirth, W. Zirwas, T. Haustein, and E. Schulz, „Implementation concepts for distributed cooperative transmission," Proc. AC- SSC '08, oct 2008.

[36] L. Thiele, T. Wirth, T. Haustein, V. Jungnickel, E. Schulz, , and W. Zirwas, „A unified feedback scheme for distributed interference management in cellular systems: Benefits and challenges for real-time implementation," Proc. EUSIPCO'09, 2009.

[37] M. Schellmann, L. Thiele, T. Haustein, and V. Jungnickel, „Spatial transmission mode switching in multi-user MIMO-OFDM systems with user fairness," IEEE Trans. Veh. Technol., vol. 59, no. 1, pp. 235–247, Jan. 2010.

[38] V. Jungnickel, T. Hindelang, T. Haustein, and W. Zirwas. SC-FDMA waveform design, performance, power dynamics and evoluation to MIMO. In IEEE International Conference on Portable Information Devices. Orlando, Florida, March 2007.

[39] V. Jungnickel, V. Krueger, G. Istoc, T. Haustein, and C. von Helmolt, „A MIMO system with reciprocal transceivers for the time-division duplex mode," in Proc. IEEE Antennas and Propagation Society Symposium, June 2004, vol. 2, pp. 1267–1270.

[40] T. Wirth, V. Venkatkumar, T. Haustein, E. Schulz, and R. Halfmann, „LTE-Advanced relaying for outdoor range extension," in VTC2009-Fall, Anchorage, USA, Sep. 2009.

[41] FP7 European Project 318555 5G NOW (5th Generation Non-Orthogonal Waveforms for Asynchronous Signalling) 2012. [Online]. Available: http://www.5gnow.eu/

[42] G. Wunder et al., "5GNOW: Non-orthogonal, asynchronous waveforms for future mobile applications," IEEE Commun. Mag., vol. 52, no. 2, pp. 97_105, Feb. 2014.

[43] R. L. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblaetter, and U. Tuerke, „Toward energy-efficient 5G wireless communications technologies: Tools for decoupling the scaling of networks from the growth of operating power," IEEE Signal Process. Mag., vol. 31, no. 6, pp. 24– 34, Nov. 2014.

[44] 5G Automotive Association http://5gaa.org/

8

International Data Spaces

Reference architecture for the digitization of industries

Prof. D. Eng. Boris Otto
Fraunhofer Institute for Software and Systems Engineering ISST
Prof. Dr. Michael ten Hompel
Fraunhofer Institute for Material Flow and Logistics IML
Prof. Dr. Stefan Wrobel, Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

Summary and further research needs

The International Data Space (IDS) offers an information technology architecture for safeguarding data sovereignty within the corporate ecosystem. It provides a virtual space for data where data remains with the data owner until it is needed by a trusted business partner. When the data is shared, terms of use can be linked to the data itself.

Analysis of six use cases from the first phase of the prototype implementation of the IDS architecture shows that the focus lies on the standardized interface, the information model for describing data assets, and the connector component. Further use cases are planned for the next wave of implementation that are based on the broker functionality and require the use of vocabularies for simple data integration.

In addition, companies need to standardize the principles that are translated into the terms of use. These principles need to be shaped, described, documented, and implemented in a simple and understandable way. They also need to be understood in the same way by different actors in the corporate ecosystem, thus requiring semantic standardization.

Furthermore, the IDS Reference Architecture Model needs to be set in context with respect to related models. In the F3 use case, an OPC UA adapter is used. Additional use cases for integration with the Plattform Industrie 4.0 administration shell and Industrial Internet Reference Architecture are pending.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_8

The IDS Architecture is also increasingly being utilized in so-called verticalization initiatives, in healthcare and in the energy sector for example. These kinds of initiatives – like the Materials Data Space – demonstrate the cross-domain applicability of the architectural components and provide information about further development needs.

Finally, in anticipation of the future development of the use cases and utilization of the IDS, work on the economic valuation of data and on the settlement and pricing of data transactions must be accelerated.

8.1 Introduction: digitization of industry and the role of data

The digitization of society and economy is the major trend of our time. The introduction of consumer technologies into businesses, the shift towards digital services for companies' business offerings, and the networking of things, people and data open up new business opportunities in nearly all sectors of the economy. Many of these have already been analyzed by the Smart Service Welt ("Smart Service World") working group within Plattform Industrie 4.0. There are five distinguishing features of the smart service world [1].

- End-to-end customer process: smart services do not address particular customer needs but rather support an entire customer process. Sensor manufacturer Endress+Hauser, for example, does therefore not only offer measuring systems but supports customers in all of their questions from planning a production facility, from advice on product selection, installation and commissioning, and configuration during operation, to advice on replacement investments and plant replacement [4].
- Hybrid service bundles: are no longer just the products central to the product offering but bundles of physical products and digital services. An example would be the sportswear manufacturer adidas, which, in addition to running shoes, also offers the "runtastic" digital service in order to support its customers' running experience – the customer process – in its entirety.
- Data at the center: hybrid service bundles, which support the end-to-end customer process, are only possible, if companies efficiently and effectively link their own data with the data of business partners and contextual data – maintaining data sovereignty above all. This contextual data often stems from public data sources and is available as open data. Examples include location data, traffic information and so on.

- Business ecosystems: the need to combine data from various sources in order to support the end-to-end customer process leads to the development of business ecosystems. These are networks of different actors, which form dynamically around an end-to-end customer process [5]. Well-known examples are the Apple App Store and Uber.
- Digital platforms: the interaction within the ecosystem and the sharing of a broad variety of data in the interest of the end-to-end customer process require new information systems. Digital platforms provide this functionality according to the "as a service" principle.

The overall characteristics of the smart service world facilitate new business models, both for new and established companies. For the latter, the key element in digitization is not simply copying new digital business models, but rather supplementing existing strengths – often in product design, production or sales – with smart service world opportunities. One example is preventative maintenance services by plant manufacturers that draw on usage data from a number of plant operators for the early identification of maintenance needs, thus improving the overall equipment efficiency (OEE).

For the manufacturing industry in particular, digitization changes not only the product offering but also the product manufacturing side. This is because the consequence of the hybrid service bundle on the offering side is an increase in product and above all process complexity on the product manufacturing side. Industry 4.0 offers a solution to managing this complexity founded on the organizational princi-

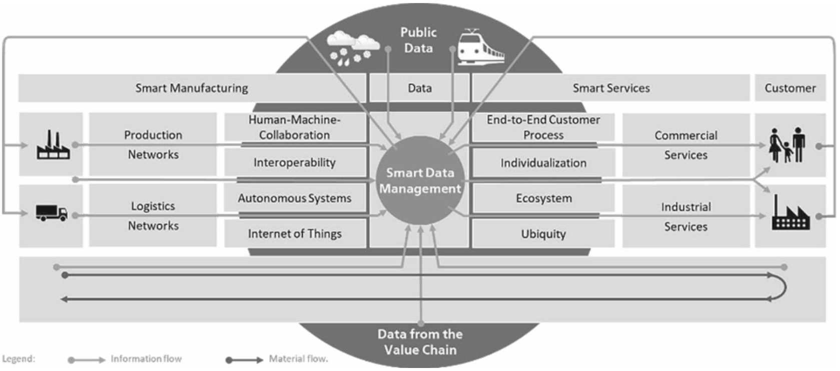


Fig. 8.1 Smart Data Management from the perspective of the focal business (Fraunhofer-Gesellschaft/Industrial Data Space)

ples of autonomy, networking, information transparency and assistance capability of systems for product manufacturing[3][2].

Both the product manufacturing and the product offering sides are linked via a modern data management system (see Fig. 8.1).

The following basic assumptions are applicable when designing this smart data management system:

- Data is a strategic and economically valuable resource.
- Companies need to share data more intensively than in the past. This applies to the efficiency, effectiveness, and especially flexibility of conventional data sharing as well as to the extent of the data to be shared. More and more, companies need to incorporate data into the business ecosystem that was previously classified as too sensitive to be made accessible to third parties.
- Companies only enter into this data sharing, if their data sovereignty is guaranteed, that is, if they can decide who is allowed to use their data on which conditions and for what purpose when the data leaves the company.

Data sovereignty is thus the capacity of exclusive self-determination of a natural person or corporate entity with respect to the asset of data [7]. Data sovereignty is expressed in the balance between the need to protect data and its shared use in the business ecosystem. It is a key competency for success in the data economy.

Alongside economic and legal frameworks, it is essential to create the information technology conditions so that data sovereignty can be exercised. This is why the IDS initiative was brought into being.

8.2 International Data Spaces

8.2.1 Requirements and aims

The IDS initiative pursues the goal of digital sovereignty in the business ecosystem. The IDS participants' aim is to be able to share data with business partners interoperably while always retaining self-determination with regards to these data assets.

This aim is to be achieved by designing an IT architecture, demonstrating its applicability and usefulness in use-case projects. The initiative is founded on two pillars: a Fraunhofer research project and the International Data Spaces Association (IDSA). The research project is being carried out by Fraunhofer funded by the Federal Ministry of Education and Research. Both the project and the association

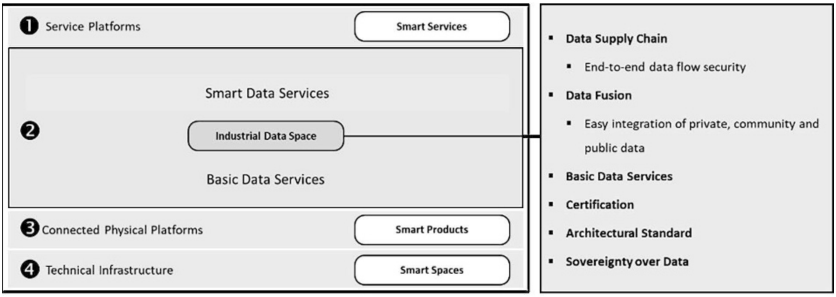


Fig. 8.2 Smart service world requirements for the International Data Space (Smart Service World working group)

are precompetitive and non-profit. The initiative encourages the broadest possible dissemination and thus also welcomes commercial paths of exploitation, which are open to all market participants.

From the overarching goal of facilitating confident data sharing in business ecosystems (smart service world), we are able to develop key requirements for the architecture of the IDS (see Fig. 8.2) [8]:

A1 Terms of use for data: when the data is shared, the data owner is able to link enforceable terms of use to the data – rules – that specify on which conditions the data may be used, by whom, and for what purpose.

A2 Secure data supply chains: the entire data value chain, from the creation/generation of the data (e.g. via sensors) to its use (e.g. in smart services) must be able to be secured.

A3 Simple data linking: in-house data must be able to be easily linked with that of business partners but also with (public) contextual data.

A4 Decentralized data storage: the architecture must not necessarily require central data storage[1]. Rather, it must be possible to only share the data with trusted partners if it is required from a clearly identifiable partner in keeping with the terms of use.

A5 Multiple operating environments: the software components of the IDS Architecture, which facilitate participation in this virtual data space need to be able to be run in conventional company IT environments, but also, for example, on IoT cloud platforms and mobile devices or sensor PCs.

---

[1] Two thirds of companies mistrust central *data lake* architectures, for example, because they fear that third parties will have unwanted access to the data cf.[10].

A6 Standardized interface: data sharing in the IDS must take place in accordance with a pre-defined and open information model.

A7 Certification: software components and participants must be certified with respect to keeping to the requirements of Industrial Data Space software and its operation. The certification criteria are the responsibility of the International Data Spaces Association.

A8 Data apps and app store: data services (data apps) provide essential functions based on the reference architecture for handling data in the IDS. Examples are data format conversion and assigning terms of use to the data. These data services need to be made available via an app store functionality.

These requirements guide the development of the Reference Architecture Model in the IDS.

8.2.2 International Data Space Reference Architecture Model

The Reference Architecture Model defines the International Data Space's individual components and their interaction [9] in terminologically and conceptually consistent terms. It distinguishes five levels and three perspectives.

The levels are:

- The business level describes the roles in the IDS.
- The functional level describes, in technological- and application-agnostic terms, the specialist and functional requirements for the IDS.
- The process level describes the interactions between the roles and the associated specialist functions.
- The information level describes the entities within the IDS and their relationships to one another in domain-independent terms.
- The system level describes the software components of the IDS.

The three perspectives are *security*, *certification* and *data governance*. They run perpendicular to the levels.

Fig. 8.3 depicts the model of roles in the IDS as part of the business level.

Data Owner, Data Provider and Data User are the core roles in the IDS. Data sharing between these core roles is supported by the following intermediary roles:

- Broker: enables data sources to be published and found.

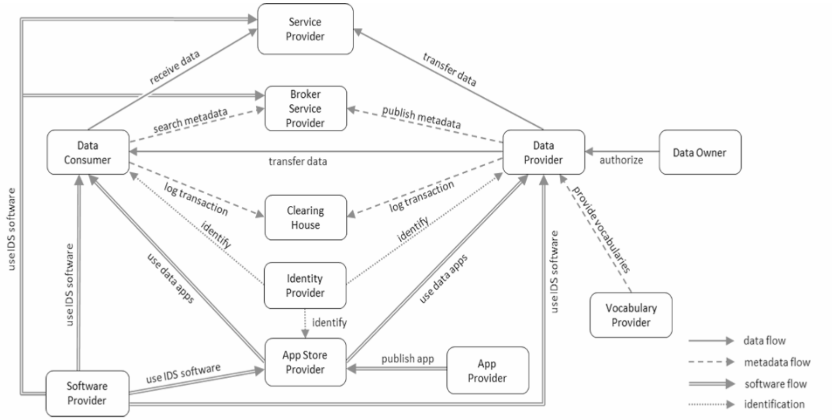


Fig. 8.3 Model of roles in the IDS (Fraunhofer-Gesellschaft/International Data Spaces Association)

- Clearing house: logs data sharing processes and resolves conflicts.
- Identity provider: issues digital identities/certificates.
- Vocabulary provider: provides semantic information models to, for example, specific domains[2].

Software and services providers offer the software and services necessary for fulfilling the different roles.

Finally, the certification authority and one or more inspection bodies ensure that all Industrial Data space requirements are fulfilled by the software and the participants.

8.2.3 State of development

The work is organized as a consortium-based research project[3] with researchers from a total of twelve Fraunhofer institutes working alongside representatives of the companies within the International Data Spaces Association. The project follows a design-based research approach where the Reference Architecture Model is implemented in software prototypes in agile development sprints. The companies use the

---

[2] The IDS draws on VoCol technologies, software that supports the shared creation and management of vocabularies [9].

[3] For details on consortium-based research cf. [6].

prototypes in their use cases in order to demonstrate the feasibility and applicability of the architecture and the associated benefits, as well as identifying additional development needs.

Twenty months into the project, six software components are available as prototypes:

K1 Connector: The central component is the IDS Connector that is used by the core roles. Prototypes are available for a basic version (without functionality for exercising usage control), for a so-called Trusted Connector (with usage control), for a sensor connector, and for an embedded version (for mobile end devices, too).

K2 Usage control: A version of the INDUCE framework [11] and processes for data labeling are available as technology packages and are additionally integrated into the Trusted Connector.

Table 8.1 Quality function illustration for the IDS

		K1				**K2**	**K3**	**K4**	**K5**	**K6**
		Connecto								
		Base Connector	**Trusted Connector**	**Sensor Connector**	**Embedded Connector**	**Usage Control**	**Information Model**	**App Store**	**Base Broker**	**Identity Provider**
A1	Terms of use for data		X	(X)	(X)	X	X			
A2	Secure data supply chain			X			X			X
A3	Simple data linking						X		X	
A4	Decentralized data storage	X	X	X	X		X		X	
A5	Multiple operating environments	X	X	X	X					
A6	Standardized interface	X	X	X	X		X	(X)	X	
A7	Certification						X			
A8	Data apps and app store							X		

Key: X – Requirement fulfilled; (X) – requirement partially fulfilled

K3	Information model: The model has not only been conceptualized, it is also available as a software library and can thus be used by software developers within the International Data Spaces Association.
K4	App store: An initial prototype is available.
K5	Base broker: An initial prototype of the broker with basic registry functionality is available.
K6	Identity provider: An initial prototype version of an identity provider service is available.

Table 8.1 shows to what extent the requirements for the IDS are fulfilled by six software components.

All of the requirements are fulfilled by at least one component. Thus, the requirement for the terms of use is addressed both conceptually in the information model and implemented in initial versions, in particular as the Trusted Connector. The requirement for various operating environments is comprehensively implemented via four versions of the Connector architecture. However, requirements A7 (Certification) and A8 (Data apps and app store) are only addressed in initial basic components.

8.3 Case studies on the International Data Space

8.3.1 Collaborative supply chain management in the automotive industry

A growing number of models, derivatives, in-car functions, and shorter product lifecycles are leading to increasing supply chain complexity in the automotive industry. Modern mid-size vehicles offer so many options and versions that theoretically more than 10^{30} varieties of configuration are possible. Around three quarters of the components required for this do not originate from the manufacturer (original equipment manufacturers, OEMs) but from suppliers. This product complexity can only be efficiently and effectively replicated in production and logistics processes, if suppliers and OEMs work closely together, not only at the planning stage (of requirements, capacities and job sequencing) but also during the execution of the processes (during transport, manufacture, and assembly). In addition to data that has been shared for decades through electronic data interchange (EDI) solutions, more and more data must be shared today that was considered too sensitive to be shared in the past. This includes:

- Inventory days for specific critical components
- Detailed information on manufacturing steps for critical components in the supplier network
- Structure of the supplier network
- Added value information for component transport (heat, shocks and vibrations, etc.)

This data will only be shared, if the data owner can specify the terms of use. This is where the IDS comes into play.

Fig. 8.4 shows the software systems and data flows in the use case of collaborative supply chain management in the automotive industry. In the first phase, the use case encompasses a tier 1 supplier and an OEM. Both use the Base Connector to support the following data sharing operations during the project's initial phase:

In step one, the tier 1 supplier informs the OEM of a supply risk for one of his suppliers (tier 2) for a subcomponent that the OEM needs. To do this, the tier 1 supplier combines data from his risk management system with master records from the supplier system and transmits this data to the OEM via the IDS Connector, including the terms of use that this data is only to be used for a specific supplier management purpose for a period of 14 days.

The OEM imports this data into its supplier management system and uses it to calculate the updated inventory days for specific components that it obtains from the tier 1 supplier. The OEM in turn sends this inventory days data via the IDS

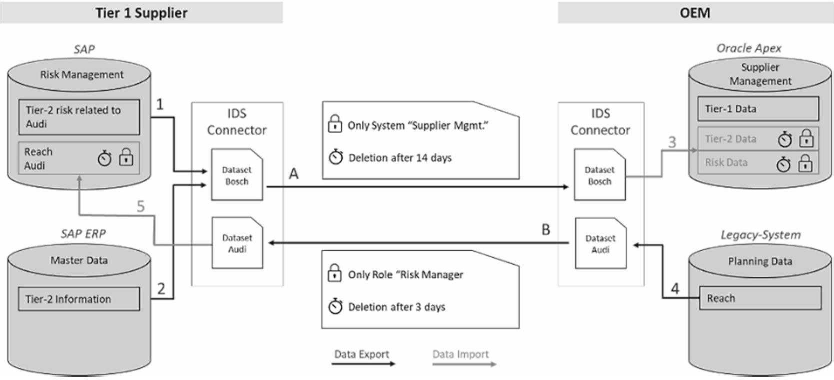


Fig. 8.4 System components in collaborative supply chain management (Chair for Industrial Information Management TU Dortmund University - formerly Audi Endowed Chair Supply Net Order Management TU Dortmund University)

Connector to the tier 1 supplier, again including the terms of use, stating the purpose for use (risk management) and the maximum days of use (three days).

The benefit for the OEM of this use case is obvious. Supply risks in production, "breaks in the supply chain", are recognized sooner and manufacturing downtime therefore avoided. For the tier 1 suppliers, the benefit lies principally in improved ability to plan their own production since the OEM's inventory days are made available to them.

This use case is currently in the software prototype implementation phase.

8.3.2 Transparency in steel industry supply chains

Steel production is a transport intensive business, with individual truck shipments being subject to interruption generally due to delays during transportation (traffic on the main leg on the freeway, situation at the factory gate, etc.).

All of the stakeholders in the supplier network – in particular the supplier and the logistics service provider as well as the individual transportation company, the producing company, additional logistics service providers for distributing the fin-

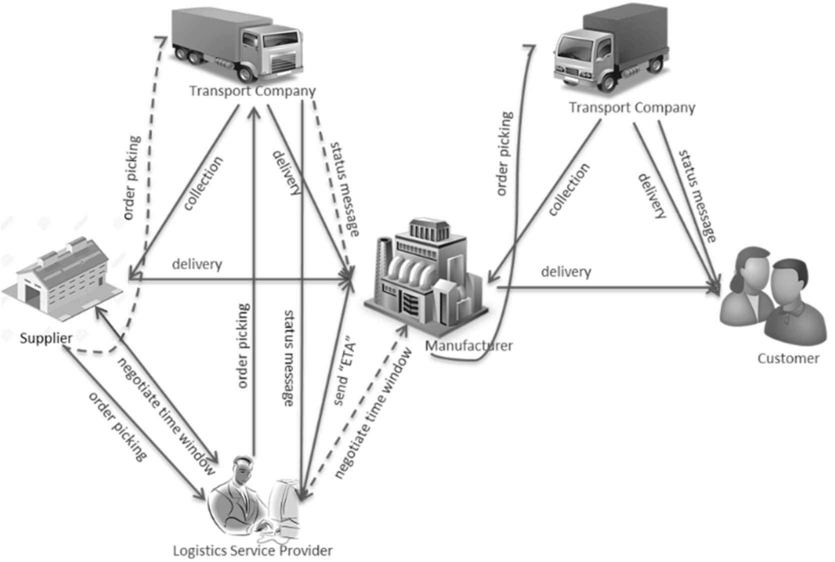


Fig. 8.5 Transparency in steel industry supply chains (Fraunhofer ISST)

ished products and the end customer themselves – have an interest in being informed in real time about events in the supply chain that lead to plan changes.

This use case addresses notification of delayed arrival on the inbound side of the manufacturing company. Here, the transportation company informs the logistics service provider, using a mobile version of the IDS Connector, that a specific shipment is delayed and states the reason for this delay. The logistics service provider calculates an updated expected arrival time and transmits this to the manufacturing company awaiting the consignment. The manufacturing company, in turn, confirms the new arrival time and transmits updated details in regards to the unloading point.

Data is shared according to the IDS information model. For the payload itself "packed" in the IDS notification, GS1 EDI XML is used.

The benefit for the manufacturing business lies in the improvement of the planning of inbound logistics processes/yard management (including staff planning within the receiving department) and production (including job controlling). For the logistics service provider, the use case offers value-added services for customers. For the transportation company the check-in process at the manufacturing company is made easier because updated information regarding time slots and unloading points is always available.

This use case is currently being piloted with thyssenkrupp and already covers more than 500 shipments per month.

8.3.3 Data trusteeship for industrial data

The model of roles for the IDS is fundamentally structured so as to allow for a separation of the role of Data Owner from that of Data Provider. This allows for new business models, specifically data trusteeship.

The processing of data across different service and process steps involves significant requirements in terms of data protection and data sovereignty that not all companies render themselves, but obtain as a service. The data trustee thus needs to ensure that data does not leak out to competitor companies, that personal data is adequately anonymized, and that services are instructed to only use data in accordance with the terms, deleting the data where necessary after use.

Independent audit firms and technical testing organizations are well suited to provide data sovereignty services in business ecosystems, including the following:

- Reviewing rules on terms of use for any conflicts
- Monitoring observation of terms of use
- Monitoring data transactions as a clearing house

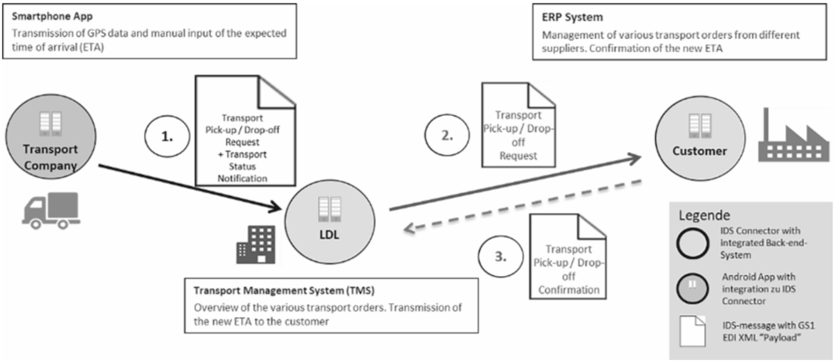


Fig. 8.6 System chain for transparency in the supply chain (Fraunhofer ISST)

- Data processing, enrichment, and provision on behalf of third parties
- Auditing and certification services within the IDS

The certification requirements and criteria, roles and certification levels, and auditing methods in the IDS are defined in a certification schema. Every organization that participates in the IDS is audited and certified according to this schema, as are all of the IDS's software components.

In addition to this, the certification body also facilitates, for example certificate-based identity management within the IDS, which is the technical basis for all Connector implementations.

This use case is currently in the conception phase.

8.3.4 Digital networking of manufacturing lines

Industry 4.0 is an organizational principle for the industrial operations of the future and rests on, among other things, the networking of all resources within product manufacturing. Machines, facilities, and staff within manufacturing are able to share information in near real time, transmitting manufacturing status data and order data, but also receiving contextual information on individual production steps. Since in many industries today manufacturing happens in distributed production networks, companies have two challenges to overcome:

- Shared semantic descriptions of manufacturing resources in the production network
- Confident data sharing between the individual resources (e.g. machines)

The use case in the IDS addresses both challenges.

Linked data principles such as RDF as the lingua franca for data integration and the associated W3C standards form the basis for the semantic descriptions. In this way, information architectures that have evolved over time can be transformed into knowledge-based information networks that then form the common informational basis for digital product manufacturing processes and innovative smart services. The data here is structured and semantically enriched in such a way that existing data silos are overcome, and data value chains can be established across functions and processes. One of the use cases for this in the IDS is the development of a knowledge graph for the production of the company Schaeffler. Here, concepts such as the administration shell model from the Industry 4.0 reference architecture model are used alongside the IDS information model. For the connection, data from the source systems (e.g. manufacturing execution systems and sensor data from machines and production) is transformed into RDF vocabularies.

If a shared semantic model – a vocabulary – is established in the production network then data can be confidently shared and its meaning also immediately be understood.. The digitization of manufacturing lines use case also provides an OPC-UA adapter that can be run as a data service in the Industrial Data Connector; it facilitates the importing of OPC UA compliant data and its linking with data from other sources (e.g. ERP systems, see Fig. 8.7). In this way, the data is made available for a range of usage scenarios in the production network.

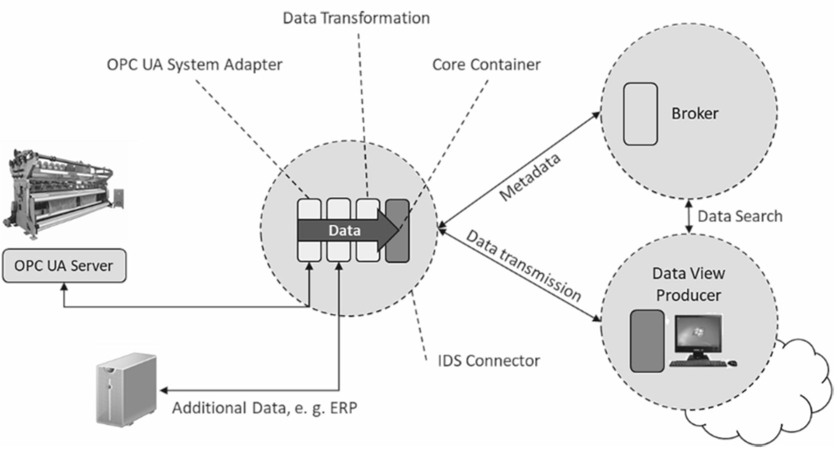


Fig. 8.7 Digital networking of manufacturing lines (Fraunhofer IOSB)

8.3.5 Product lifecycle management in the business ecosystem

More than two thirds of all new products are based on new materials. In order to safeguard the innovational strength and retain technological sovereignty and closed value chains within Germany, universal digitization of material and product properties over their entire lifecycle ("from ore to the fridge") is of strategic importance.

The Materials Data Space provides digitized information on materials and their properties, components, and their alterations in manufacturing and use over the entire value chain, thus covering their entire lifecycle from use to strategic recycling. The Materials Data Space, a verticalization of the IDS architecture, is a strategic initiative of the Fraunhofer MATERIALS group, which aims for comprehensive digital image within the entire business ecosystem (see Fig. 8.8).

The steel industry, for example, does not only sell steel strips themselves but also the digital image of the steel strip (microstructure, composition, inclusions, material history, etc.). The combination of the physical product together with the digital image is a key success factor for the future competitiveness of companies such as Salzgitter AG, the implementation partner in this use case. In addition, combining the data over the entire lifecycle results in shorter development times and learning

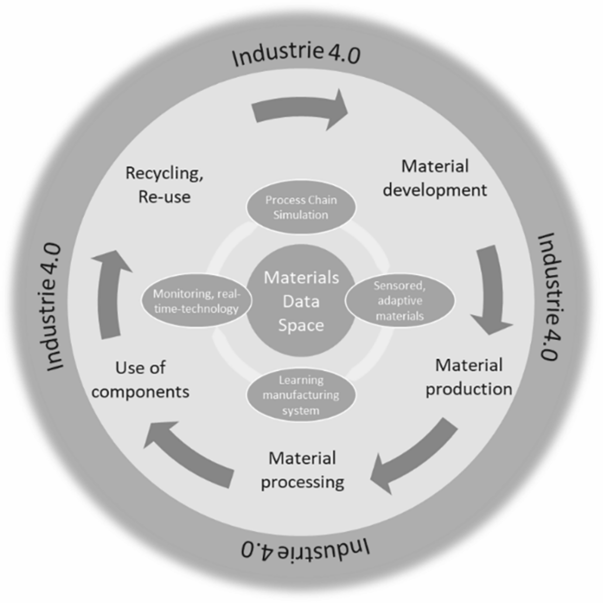


Fig. 8.8 Materials Data Space (Fraunhofer-Gesellschaft)

manufacturing processes. What is more, fundamental potential benefits with respect to materials and production efficiency and recycling accrue for the entire business ecosystem.

This use case is supported by a Fraunhofer preliminary research project identifying domain-specific requirements for IDS components as well as for the terms of use that will be linked with the data of the various actors in the Materials Data Space business ecosystem.

8.3.6 Agile networking in value chains

Networking in value chains is not only limited to companies or people but, with the spread of the Internet of Things, increasingly stretches to objects, too. Load carriers, containers, trucks, etc. are all uniquely identifiable, communicate with their environment, and make decisions independently (with respect to shipping routes, for example). In the course of this, these objects produce data, specifically value chain status data. They provide information, for example, on locations (including time stamps), atmospheric conditions (temperature, humidity), and shocks.

Agile networking in value chains requires both confident sharing of this kind of status data by the objects themselves and the identification of relevant data sources.

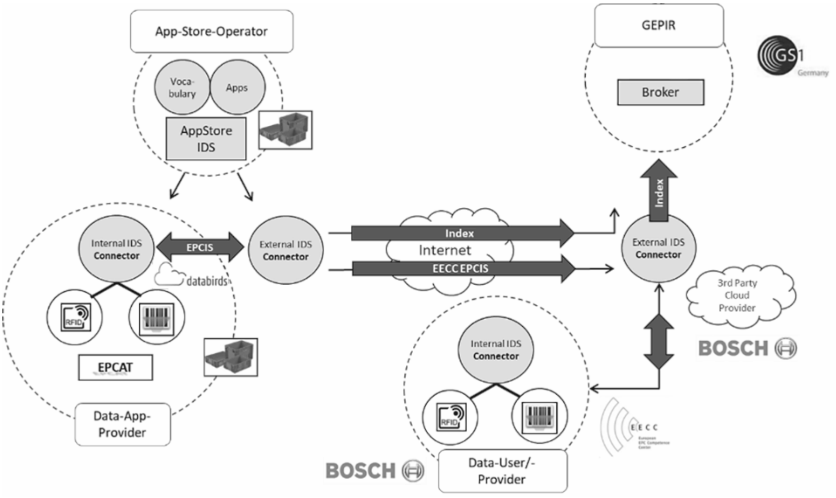


Fig. 8.9 Overall architecture for the STRIKE use case (Fraunhofer-Gesellschaft/International Data Space)

This is because master records and status data from small load carriers are shared in value chains with changing business partners across companies.

In the STRIKE use case (Standardized RTI Tracking Communication Across Industrial Enterprises), status data is sent via the IDS Connector in a useable format for the recipient. This status data is produced and managed using the EPCIS standard. Using the example of new load carriers equipped with RFID, a demonstration is given of how the status data can be assigned to the data owner within the GS1 community. The data owners therefore remain sovereign over their data.

The existing GS1 partner registry, GEPIR (Global Electronic Party Information Registry), is also being developed into an Industrial Data Space broker in order to be able to identify providers of status data in the value chain. In addition, this example shows how value-added services can be derived from the data with the help of apps.

Fig. 8.9 shows the overall architecture for the STRIKE use case.

8.4 Case study assessment

Use cases assessment mirrors the strategy for the implementation of the prototype reference architecture model (see Table 8.2). This is because prototypes of the IDS Connector, as the architecture's core component in various implementations, are being put to work or have been planned in all of the use cases. In addition, the use cases confirm the coexistence of various Connector implementations with different functional characteristics that nevertheless all adhere to the principles of the reference architecture model.

The prerequisite here is conformity to the IDS information model. This criterion is being fulfilled in all cases.

This analysis also shows the step-by-step development or implementation of important functionalities, since usage control technologies are neither in use nor planned for all cases.

In addition, the need of action with respect to implementation and use of the app store and broker components becomes apparent.

Table 8.3 additionally shows to what extent the strategic requirements for the IDS architecture are a reality in the six use cases.

This analysis demonstrates that the focus in the use cases during the initial phases of the initiative lies on the standardized interface, which is in use in all of the examples. The requirement for decentralized data storage is also being fulfilled in nearly all of the use cases.

Table 8.2 Use of software prototypes in the use cases

		K1				K2	K3	K4	K5	K6
		Connector								
		Base Connector	Trusted Connector	Sensor Connector	Embedded Connector	Usage Control	Information Model	App Store	Base Broker	Identity Provider
F1	Collaborative SCM	X	(X)			X	X			
F2	Transparency in supply chains	X		X	X		X		(X)	
F3	Data trusteeship for industrial data		X			X	(X)		(X)	X
F4	Digital networking of manufacturing lines			X		(X)	X		X	
F5	Product lifecycle management in the business ecosystem	X				(X)	X		(X)	
F6	Agile networking in value chains	X				(X)	X	(X)	(X)	(X)

Key: X – prototype in use; (X) – use planned.

Specific requirements – such as those for a secure data supply chain and multiple operating environments – feature in those use cases where these specifications are relevant.

Areas that have been addressed only to a limited extent in the project's initial phase are the terms of use (only currently being utilized in F1) and the possibility to provide data apps. The latter is no surprise since an app ecosystem can only develop with the increasing dissemination of the initiative. The same applies to data linking: this requirement will increasingly move to the fore of implementation as use case scenarios become more complex.

The requirement to be able to attach terms of use to the data will become increasingly significant during future use case implementation phases when the basic communication is able to be conducted via the IDS and when increasingly sensitive data elements are being shared.

Table 8.3 Fulfillment of requirements in the use cases

		K1	K2	K3	K4	K5	K6
		Collaborative SCM	Transparency in supply chains	Data trusteeship for industrial data	Digital networking of production lines	Product lifecycle management in the business ecosystem	Agile networking in the value chain
A1	Terms of use for data	X		(X)		(X)	(X)
A2	Secure data supply chain		X				X
A3	Simple data linking	(X)	(X)		X	X	(X)
A4	Decentralized data storage	X	X	X		X	X
A5	Multiple operating environments		X		X		(X)
A6	Standardized interface	X	X	X	X	X	X
A7	Certification			X			
A8	Data apps and app store	(X)			X		(X)

Key: X – prototype in use; (X) – use planned.

Sources and literature

[1] Arbeitskreis Smart Service Welt. (2015). Smart Service Welt: Umsetzungsempfehlungen für das Zukunftsprojekt Internetbasierte Dienste für die Wirtschaft. Berlin.

[2] Bauernhansl, T., ten Hompel, M., & Vogel-Heuser, B. (2014). Industrie 4.0 in Produktion, Automatisierung und Logistik. Anwendung · Technologien · Migration. Berlin: Springer.

[3] Hermann, M., Pentek, T., & Otto, B. (2016). Design Principles for Industrie 4.0 Scenarios. 49th Hawaii International Conference on System Sciences (HICSS 2016) (S. 3928-3927). Koloa, HI, USA: IEEE.

[4] Kagermann, H., & Österle, H. (2007). Geschäftsmodelle 2010: Wie CEOs Unternehmen transformieren. Frankfurt: Frankfurter Allgemeine Buch.

[5] Moore, J. F. (2006). Business ecosystems and the view from the firm. Antitrust Bulletin, 51(1), S. 31-75.

[6] Österle, H., & Otto, B. (Oktober 2010). Konsortialforschung: Eine Methode für die Zusammenarbeit von Forschung und Praxis in der gestaltungsorientierten Wirtschaftsinformatikforschung. WIRTSCHAFTSINFORMATIK, 52(5), S. 273-285.

[7] Otto, B. (2016). Digitale Souveränität: Beitrag des Industrial Data Space. München: Fraunhofer.

[8] Otto, B., Auer, S., Cirullies, J., Jürjens, J., Menz, N., Schon, J., & Wenzel, S. (2016). Industrial Data Space: Digitale Souveränität über Daten. München, Berlin: Fraunhofer-Gesellschaft; Industrial Data Space e.V.

[9] Otto, B., Lohmann, S. et al. (2017). Reference Architecture Model for the Industrial Data Space. München, Berlin: Fraunhofer-Gesellschaft, Industrial Data Space e.V.

[10] PricewaterhouseCoopers GmbH. (2017). Datenaustausch als wesentlicher Bestandteil der Digitalisierung. Düsseldorf.

[11] Steinebach, B., Krempel, E., Jung, C., Hoffmann, M. (2016). Datenschutz und Datenanalyse: Herausforderungen und Lösungsansätze, DuD – Datenschutz und Datensicherheit 7/2016

EMOIO Research Project

9

An interface to the world of computers

Prof. D. Eng. Hon. Prof. Wilhelm Bauer ·
Dr. rer. nat. Mathias Vukelić
Fraunhofer Institute for Industrial Engineering IAO

Summary

Adaptive assistance systems are able to support the user in a wide range of different situations. These systems take external information and attempt to deduce user intentions from the context of use, without requiring or allowing direct feedback from the user. For this reason, it remains unclear whether the system's behavior was in accordance with the user's intentions – leading to problems in the interaction between human and adaptive technology. The goal of the EMOIO project is to overcome potential barriers of use with the aid of neuroscientific methods. Merging ergonomics with the neurosciences into the new field of neuroergonomics research produces enormous potential for innovation, to make the symbiosis between humans and technology more intuitive. To this end, brain-computer interfaces (BCIs) offer a new generation of interfaces between humans and technology. BCIs make it possible to register mental states such as attention and emotions and transmit this information directly to a technological system. So-called neuroadaptive systems continuously use this information in order to adjust the behavior, functions or the content of an interactive system accordingly. A neuroadaptive system is being developed by a consortium of partners from research and industry as part of the EMOIO project. The goal of the system is to recognize, based on the users' brain activity, whether system-initiated behaviors are approved or rejected. The system is able to use this information to provide the person with the best possible assistance and thus adapt to individual and situational demands. To do this, neuroscientific methods such as electroencephalography (EEG) and

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_9

functional near-infrared spectroscopy (fNIRS) are being evaluated with respect to their suitability for measuring emotions (approval/rejection).

In addition, a corresponding algorithm is being developed for real-time emotional recognition. The miniaturization and resilience of the EEG and fNIRS sensors are also being promoted. Finally, the developed system is being explored in three different areas of application: web-based adaptive user interfaces, vehicle interaction, and human-robot collaboration.

Project outline data

Project aim

The aim of the project is to use neuroscientific methods to reduce barriers to the use of assistance systems. To do this, Fraunhofer IAO, together with partners from research and industry, is developing a neuroadaptive system as part of the EMOIO project. The system is intended to recognize, based on the users' brain activity, whether the user approves or rejects the system's behavior, and adapt that behavior accordingly.

Cooperation partners

University of Tübingen – Institute of Medical Psychology and Behavioral Neurobiology at the Faculty of Medicine and Institute of Psychology, NIRx Medizintechnik GmbH, Brain Products GmbH, University of Stuttgart – Institute of Human Factors and Technology Management (IAT)

Research plan

Project schedule:

Phase 1: 01/2015 to 12/2016
Phase 2: 01/2017 to 12/2017

Key findings

- Representative neuronal correlates of affective responses during interaction with technology
- Real-time classification of neuronal correlates of affective user reactions
- Miniaturization of EEG/fNIRS hardware sensors and optimized simultaneous EEG/NIRS measurement
- Testing of real-time classification in three fields of application: smart home, human-robot collaboration, vehicle interaction

Contact

Dr. rer. nat. Mathias Vukelić,
mathias.vukelic@iao.fraunhofer.de

9.1 Introduction: designing the technology of the future

Digitization is having a vital impact on human's world and everyday lives, and interaction with digital technology is becoming ever more self-evident and important. Increasingly intelligent technological products are finding their way into our everyday working lives. Digitization fosters the networking of various technological systems, facilitating greater ease of communication and cooperation with one another as we use them in our everyday and working lives. Autonomous robots are making industrial production easier, and architects and designers are planning and developing their solutions in virtual reality. These kinds of solutions are facilitating improvements in productivity and efficiency, with corresponding time savings. Nevertheless, the increasing integration of technology in our workplaces also entails new challenges and potential for conflict. Often, humans with their individual preferences and needs find themselves overlooked in the development of technological systems. The resulting solutions, whilst technologically advanced, may nevertheless offer limited gains in terms of the productivity, creativity, and health of the users in question. The challenge facing this growing use of technology is to create suitable working environments where humans can receive the best possible support in their tasks in the broadest range of situations.

Human-technology interaction and neuroergonomics

The increasing digitization of the workplace means research findings that help us understand and improve the interaction between human and technology – and therefore facilitate humans using technology products efficiently – are of key importance. Set against this backdrop, the research field of human-technology interaction is constantly growing in importance. In addition to usability, also cognitive and emotional user factors play a key role here, especially in the workplace. This raises the following questions: how great is the human cognitive load when they are working? What happens to human's emotional well-being when they are interacting with technology? The classical psychological and engineering methods used in ergonomics provide inadequate answers to these questions. There is thus a need for supplementing existing procedures with a component that facilitates access to the user's implicit mental state. This would allow the cognitive and emotional processing, which is not immediately apparent to the conscious mind, to be made accessible to the interaction design process.

Since 2012, under the heading of neuroergonomics, Fraunhofer IAO has been researching the extent to which neuroscientific methods may be used to capture cognitive and emotional user states during technology use, and to facilitate new perspectives and technological options in ergonomics. The institute is pursuing an interdisciplinary research approach here, combining expertise from the fields of psychology, information technology, engineering and neuroscience. Neuroergonomics is considered a research field of the future with great potential for science and economic praxis: ranging from workplace design, through virtual engineering and vehicle ergonomics, to user-friendly IT systems. Appropriate techniques for recording and measuring brain activity that are non-invasive, inexpensive, and secure are electroencephalography (EEG) on the one hand and functional near-infrared spectroscopy (fNIRS) on the other. EEG directly captures the sum of the electrical activity of nerve cells in the brain by recording voltage fluctuations via electrodes on the scalp [1]. fNIRS measures the metabolic processes related to nerve cell activity, specifically the amount of oxy- and deoxygenated hemoglobin in the blood vessels and thus captures changes in blood flow in the brain [1]. These are the two techniques that are primarily used in the institute's "NeuroLab" to measure brain activity, measurement techniques, which are also well suited to mobile application scenarios. Based on the brain's activation patterns, it is thus possible to measure and quantify various cognitive and emotional states of the user that are relevant in the working context. Against the backdrop of the increasing significance of emotional aspects in human-technology interaction, an initial pilot project has already demonstrated that basic brain processes of the "user experience" [2] during technology use [3] can be measured as objectively as possible with neuroscientific techniques.

9.2 Adaptive and assistance systems

One of the key issues, that Fraunhofer IAO has been working on intensively in the context of neuroergonomics research, is adaptive and assistance systems. Ever more frequently today we encounter interactive assistance systems that are able to act independently in certain everyday situations or carry out tasks in the workplace autonomously. Assistance systems that act independently can support users in a wide variety of situations. They can be seamlessly integrated into everyday life and can help to reduce the fear that less technologically savvy users have of interacting with increasingly complex digital systems. Nevertheless, these systems are not always purely beneficial; conflict can also arise precisely during the interaction between user and adaptive technologies. In order to support users, these kinds of systems have to fall back on external information and attempt to deduce the aims and intentions of the human from the context of use. Current technological systems are insufficiently equipped to observe their users without any interruption and delay and thus draw corresponding conclusions in real time. It thus remains unclear whether the system's behavior was in the interest of the user. Instead of the intended support being provided by the system's adaptive behavior, it can lead to users feeling a loss of control and rejection towards the system. The question arises, then, how assistance systems can be designed in future so that this potential for conflict is reduced and user acceptance increased. Whereas intelligent systems already make use of a range of contextual data such as devices used, environmental conditions, and distance from the display [4–6] in order to offer optimal adaptation, the potential for emotion sensing as an input parameter for adaptive systems is as of yet largely untapped.

In the EMOIO project, sponsored by the Federal Ministry of Education and Research (*Bundesministerium für Bildung und Forschung*, BMBF), Fraunhofer IAO and its five additional partners have set themselves the goal of developing a brain-computer interface that captures the subjectively perceived suitability (approval/rejection) of system-initiated assistance behaviors, evaluates it, and delivers it to an adaptive system so that its assistance functions can be optimally adapted to the user. The following section provides information on the initial results of the project, a project which is part of the human-technology interaction and neuroergonomics research fields.

9.3 Brain-computer interface and neuro-adaptive technology

The period where we only communicated with computers using input media such as a mouse and keyboards is over. Some technological products let us talk to them while others respond to gestures. Technological products that are becoming increasingly intelligent are being integrated into everyday working life. The brain-computer interface (BCI) is currently the most direct form of an interface for interaction and communication between user and technological systems. BCI applications have thus far been concentrated first and foremost on the clinical environment. The interface enables users with physical or perceptual limitations, such as after a stroke or locked-in patients, to communicate with their environment, surf in the Web, or even paint [7–11]. Furthermore, BCIs have been used in the context of neurofeedback training for treating psychiatric conditions such as depression and schizophrenia [12–14]. Users without physical impairment, too, can benefit from this kind of interface in their everyday lives or at work. Alongside the active and intentional control of BCIs by the user, so-called "passive" BCIs have also become established in recent years [15]. Passive BCIs do not require intentional control by the individual. They capture cognitive and emotional states such as affect, mental load, or surprise effects, and transmit these directly to a technological system [16][17]. If this information is used as a basis for correspondingly adapting a system's content and functions or user interface during interaction, then it is known as a neuro-adaptive technology [19]. This kind of technology is formed essentially of three parts (see Fig. 9.1).

The first part (afference) consists in collecting the available and observable data about the user and the user context via neuroscientific measurement techniques (e.g. EEG or fNIRS). In the second stage (inference), this data is classified with the aid of machine learning algorithms, so that relevant cognitive or emotional information from the user can be interpreted. In the final stage (efference), the decision regarding the system's adaptation behavior and the execution of the adaptation on the

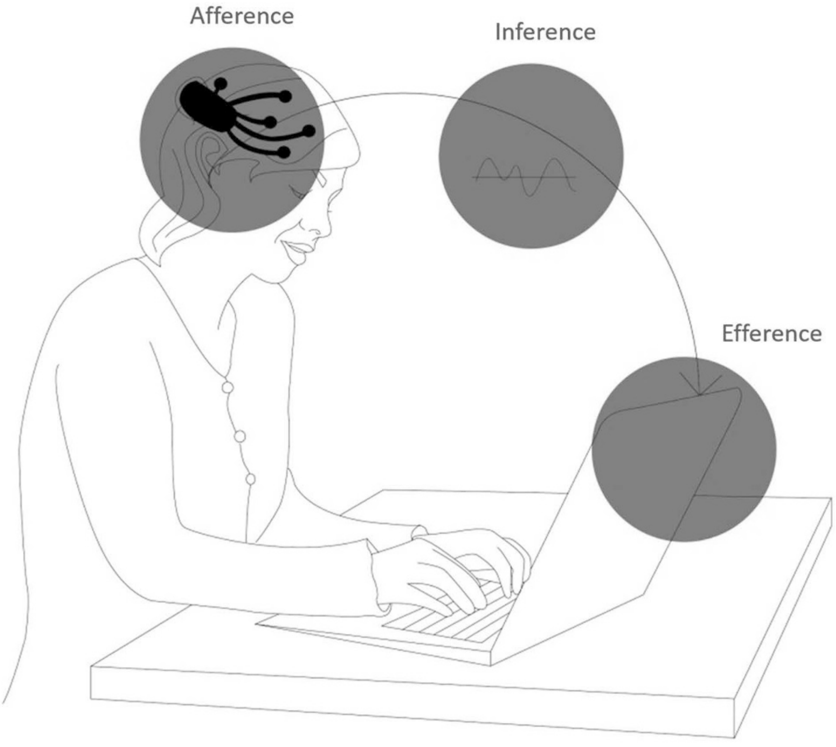


Fig. 9.1 Schematic illustration of a neuro-adaptive technology (Fraunhofer IAO)

basis of the user's cognitive and emotional states (and thus transparent feedback on system behavior) is provided to the user.

Advantages of emotion sensing for adaptive and assistance systems

In daily life, emotions help us to assess situations we experience and thus to act accordingly. Just like perception and movement, emotions are represented by neuronal patterns in the brain. These patterns can be captured using appropriate sensors and computer algorithms [20]. The key issue here is improving the interaction between user's and technology. Neuroscience shows how closely the individual's emotional system is connected with their brain. Furthermore, cognitive and memory processes influence human emotional states such as anger, joy, or affect.

In psychology, emotional intelligence is said to play an essential role in human decision-making. The theory is that social interaction and communication with other individuals works because we form a mental model of the other person's current state (see "Theory of Mind" in [21]). This enables us to identify the intended actions of the person we are interacting with and adapt our actions accordingly. This ability – the ability to recognize the emotional state of our human interaction partner – has until now been completely absent in technological systems [22]. It follows, that the interaction can be decisively improved if information about the current emotional state of the user can be provided to the technology system via a BCI. In this way, there should be the option, via the BCI, to optimally adapt the assistance system's behavior to the user. Thus, the active user feedback becomes superfluous and the interaction is uninterrupted.

The vision is to consistently orient the increasing intelligence and autonomy of technology towards people's individual needs and preferences so that neuro-adaptive technologies support and assist them as effectively as possible. Despite the many advantages, there is a still a gap between the potential benefit offered and the actual value added by this technology for applications outside of the medical field – for application-oriented human-technology interaction scenarios. The EMOIO project is laying the foundations for closing this gap between basic BCI research and practical applications of human-technology interaction.

9.4 EMOIO – From basic to applied brain research

9.4.1 Development of an interactive experimental paradigm for researching the affective user reactions towards assistance functions

Up to this point, the capturing of emotional or affective states via mobile neurophysiological methods such as EEG and fNIRS has largely remained in the domain of basic neuroscientific research. Normally, the experiments take place under strictly controlled test conditions where mostly purely receptive image, audio, or video material is used to induce corresponding emotional responses in the participants.

There are also very few studies in the field of human-technology interaction research that have examined affective responses using neurophysiological methods [23][24]. Studies with a plausible interaction scenario between an user and an adaptive assistance system have until now rarely been conducted. The main aim of EMOIO's initial phase was thus to research and identify the brain's pattern of activation, that are underlying the user's emotional state of satisfaction (positive affect) or rejection (negative affect) while they are interacting with assistance systems. As part of the project, a new, interactive experimental paradigm called AFFINDU [23] was developed to identify affective user reactions to assistance functions in realistic settings. An empirical study was conducted to research the foundations of affect detection. The study simultaneously used EEG, fNIRS, and various secondary psychological measuring techniques (such as measuring muscle activity in the face, or recording the heart rate variability). In total, the affective reactions of 60 participants (aged between 18 and 71 years) to static image stimuli (standardized stimulus material from neuroscientific basic research suitable for comparative purposes) and during the interaction with AFFINDU were recorded. Fig. 9.2 provides an insight into the work at Fraunhofer IAO's "NeuroLab" whilst preparing the experiments (attaching the measuring sensors to individuals) and conducting them.

Fig. 9.2 C shows the experiment being carried out while a participant is interacting with AFFINDU. AFFINDU is a prototype of an assistance system consisting of a graphical menu interface with 16 different menu items. The system is able to induce corresponding affective responses in the users during the interaction. A detailed description of the functionality and procedure of the experiment is provided in [23]. In summary, the interaction with AFFINDU represents a plausible usage scenario in which two key system-initiated behaviors of an adaptive assistance

Fig. 9.2 Work at the Fraunhofer IAO NeuroLab: test preparations for conducting the experiment (A, attaching the measurement techniques); measurement techniques (B, EEG and fNIRS sensors, whole head coverage); carrying out the experiment (C, participant, left, interacting with AFFINDU) (Fraunhofer IAO)

system are simulated. Participants were asked to use a keyboard to navigate through the menu while they were asked to select various target menu items. This navigational task represents a very simple user goal where AFFINDU is able to support the user appropriately (positive scenario). In the positive scenario, the user goal – that is, the desired menu item – is recognized as correct by AFFINDU and the duration of navigation to the target menu item correspondingly shortened. AFFINDU may also misidentify the user's goal and thus hinder the user in reaching their goal (negative scenario). The user's emotional assessment of the system behavior can be plotted in two independent dimensions: *valence* (positive to negative) and *arousal* (calm to aroused). Thus, the system's behavior is assessed positively if it is conducive to the user's goal (the user reaches the desired menu item more quickly), whereas it is assessed negatively if AFFINDU's behavior is not aligned with the user's goal (the user needs more time to reach the desired menu item). The results of our research show that the induction and quality of affective user reactions (positive and negative) to individual assistance functions can be successfully measured during interaction using neurophysiological measuring methods such as EEG and fNIRS [23].

9.4.2 Studying the ability to detect and discriminate user affective reactions with EEG and fNIRS

In order to be able to research the brain patterns of related to the individual's affective user reactions towards system-initiated assistance functions in a specific manner, the interaction with AFFINDU was implemented in an event-related experimental trial-procedure [23]. In this way, the brain's responses within a fixed time window of interest to a given assistance event can be identified and interpreted. Objective quantification of the event-related activity can also be carried out by analyzing the amplitudes and latencies of the time course of the EEG and fNIRS signals over time. This provides the condition for specific examinations and discriminability of neuronal correlates of emotional-affective user reactions using EEG and fNIRS. The identification and localization (including the selection of representative EEG and fNIRS positions) of reliable patterns from the individual measuring modalities is thus a focus of the EMOIO research project. These neuronal correlates serve as the basis for the development of an algorithm capable of real-time evaluation and classification of EEG/fNIRS data for affect detection.

The results from the evaluations of the EEG signal dynamics over time show that it is possible to measure neuronal correlates, which permit a reliable discrimination between supportive (positive affective response) and hindering (negative affective response) assistance behaviors after just 200msec in specific EEG positions. Additional representative EEG time periods for discriminating affective user reactions can be observed from approx. 300msec and 550msec after the system-initiated assistance function.

Due to its high temporal resolution, the electrical activity measured by the EEG can furthermore be divided into several frequency bands, so-called EEG bands that are characterized by specific rhythmic oscillations. Here, oscillatory activities can be observed from low frequencies in the range from 0.1 to above 4Hz (delta waves), 4 to 8Hz (theta waves), 8 to 13Hz (alpha waves), up to faster oscillations in the range from 13 to 30Hz(beta waves) and above 30Hz (gamma waves). Analyzing the amplitudes of the EEG's different frequency components may allow further conclusions to be drawn regarding the user's cognitive and emotional processes. Furthermore, with respect to assessing positive and negative system-initiated assistance functions, the alpha, beta, and gamma EEG frequency bands have shown to be reliable neuronal correlates of an individual's emotional-affective reactions during interaction with AFFINDU.

The advantage of EEG measurements lies in their capacity for high temporal resolution, permitting an exact chronometric allocation of cognitive and emotional processes. EEG's spatial resolution, however, is very limited and usually lies in the region of a few centimeters.

In order to investigate the relationships between the individual's emotional states and brain patterns, methods are required that offer not only a very high temporal resolution of the activity of particular brain regions but also a good spatial resolution. The requirement for a good spatial resolution is fulfilled by the fNIRS method. Using fNIRS, changes can be recorded in the blood flow of the brain nerve cells can be picked up in brain regions lying up to 3 cm below the surface of the scalp. Thus, the fNIRS method is highly suited to capture the local activity in specific regions of the cerebral cortex that are related to emotional processing. The results from the evaluations of the fNIRS data from the AFFINDU experiment show that both the frontal as well as regions in the back of the cerebral cortex respond sensitively to different adaptation behaviors. It is well known that these regions are associated with motivational aspects and the semantic meaning of emotional processing in human beings. In particular, it was shown that activity in these regions increases in the case of a positive support by AFFINDU, whereas the activity decreases in the case of a negative event provided by AFFINDU. These neuronal correlates provide additional representative neuronal signatures for user's emotional-affective reactions when interacting with technology that can be combined with the patterns found from the EEG data analysis.

Research was also carried out within the project into the differences in brain activity in different age groups. We know from basic neurobiological and psychological research that developmental changes (differentiation/dedifferentiation) in cognitive and sensory functions occur with age. Emotional research has also found that the extent of positive and negative affectivity changes in elderly people, showing that subjective assessments of negative emotions appearing to diminish with increasing age [25]. We can thus assume that older people in particular will demonstrate differences in the neuronal correlates of emotional-affective reactions. We were also successfully able to demonstrate within the project that this is the case by means of additional results from the AFFINDU experiment [26]. To do this, study participants were divided into two groups based on a median split of their age: "young" (aged between 22 and 44 years), and "old" (aged between 48 and 71 years). We were able to show that the older participants experienced AFFINDU's supportive assistance function as more positive and its impeding function as less negative as compared to the younger group of participants. Furthermore, , it could been shown that the event-related EEG amplitude responses correlate with age, with differences between the two groups especially present in the time interval starting at around 300 msec after the self-initiated assistance event. These results suggests that there are distinct cognitive strategies for emotional-affective processing in the aging brain. These results show that age-related differences in the neuronal correlates must be taken into account when developing adaptive technologies towards user's current preferences and individual needs.

These neuronal correlates of emotional-affective user reactions found in the project correspond extensively with results for temporal and spatial (EEG and fNIRS positions) EEG and fNIRS patterns known from the neuroscientific literature. The results from the first part of the project thus provide an important contribution to basic research into affect detection in human-technology interaction using neurophysiological methods.

9.5 Summary and outlook

9.5.1 Summary and outlook from the research within the EMOIO project

The results presented on the neuronal correlates provide the basis for developing an algorithm that combines and evaluates EEG and fNIRS data in real-time. This data can be used to classify user's affective reactions during the interaction with technology. To do this, further studies in signal processing and machine learning algorithms for classifying the neuronal correlates were required from EMOIO project partners. The principal advantage of the multimodal approach lies in the fact that the individual strengths of the one measurement techniques can be used to compensate for the disadvantage of the other. Thus, the limited spatial resolution of the EEG can be supplemented by the fNIRS measurement techniques, thus, facilitating the localization of brain activity. On the other hand, the higher temporal resolution of EEG can compensate the deficiencies of fNIRS in this regard. Compared with a unimodal approach, the combination of metabolic and neuroelectrical data enables us to achieve improved classification accuracies for estimating the user's emotional-affective state. Furthermore, the multimodal approach reduces the susceptibility that are prone to errors of a unimodal classification approach, which might occur due to artifacts, for example due to the temporary failure of a recording modality alone or the recording of incorrect values such as muscle artifacts in the EEG. In this way, the affect classification can be carried out exclusively by using the interference-free measurement modality. For applied research, a combination of the two techniques also offers, alongside generally more reliable recording of brain activity, the benefit of an inexpensive mobile measuring technique without any usage restrictions. Building on the results achieved in the foundational empirical studies, the head-mounted sensor technology was also miniaturized by development partners from the project, thereby improving the simultaneous monitoring of EEG and fNIRS signals.

In the project's second phase, Fraunhofer IAO investigates the feasibility and added value of real-time affect detection in three fields of application: web-based adaptive user interfaces, vehicle interaction, and human-robot collaboration. In this way, Fraunhofer IAO together with the project partner is providing an important contribution to the application of neuroscientific methods for real-time classification of emotional-affective user reactions. This lays the foundation for applying a neuro-adaptive system as a supplementary source of information for independently acting adaptive systems.

9.5.2 Outlook and applications of brain-computer interfaces

A report published recently by the BNCI Horizon 2020[1] project sponsored by the EU [26] shows that there are currently 148 BCI-related industrial stakeholders in the market. These stakeholders encompass a variety of sectors such as automotive, aerospace, medical technology, rehabilitation, robotics, and entertainment and marketing. The information and communications sector offers a large potential for the long-term development and integration of BCIs. Especially, in the fields of multimodal operating systems and ambient intelligence. Companies such as Microsoft[2] and Philips[3] have been looking into the question of how neuroscientific methods can be integrated into human-technology interaction for many years. In addition, other technology-driven companies from Silicon Valley such as Facebook[4] and Elon Musk's recently founded start-up NeuraLink[5] are also investing heavily into the development of future BCI applications. An additional field for applied research might also be the factory of the future. Cognigame[6] and ExoHand[7] are a couple of research projects by FESTO that investigate interaction concepts using BCI in the industrial context. The automotive industry, too, would see benefit from BCI's in

---

[1] http://bnci-horizon-2020.eu/

[2] http://research.microsoft.com/en-us/um/redmond/groups/cue/bci/

[3] http://www.design.philips.com/about/design/designnews/pressreleases/rationalizer.page

[4] https://www.scientificamerican.com/article/facebook-launches-moon-shot-effort-to-decode-speech-direct-from-the-brain/

[5] https://www.wsj.com/articles/elon-musk-launches-neuralink-to-connect-brains-with-computers-1490642652

[6] https://www.festo.com/cms/de_corp/12740.htm

[7] https://www.festo.com/cms/de_corp/12713.htm

the future. For example, BCI could model driver states for situational adaptive driver assistance. Thus, an intelligent system could warn the driver if he or she is tired, stressed, or distracted from information gathered via the BCI. Jaguar Land Rover[8], Nissan[9] and Daimler [28] have already demonstrated such concepts of driver modelling through various research projects. Capturing brain states via BCI also offers interesting potential applications in the field of digital media for knowledge and training software. For example, training software could adapt itself to the user's momentary receptiveness and ability to concentrate and thus controlling theamount of learning material such that the user is not overwhelmed.

In summary, we can say that neuroergonomics is still a relatively new field of research, which is why neuro-adaptive technology use is oriented around future possibilities. Whether in the end, this kind of technology really has the potential to be a success depends on user acceptance as well as technological feasibility. Especially, in the area of user acceptance there is still, room for improvement due to the limitations of current sensor technology. However, neuroscientific research is already providing promising results in terms of relevant improvements in the miniaturization and mobility of the sensor technology [29]. Furthermore, this field will be advanced quite significantly also by the entertainment industry[10]. They are working on new design concepts for head-mounted sensor technology that can be used in a broad range of applications. Neuro-adaptive technologies offer a huge potential for use in the workplace. It is precisely in this context that diverse questions arise, ranging from issues of user's individual autonomy to data privacy. This means that neuro-adaptive technologies are also the subject of ethical discussions. These require future academic debate by incorporating so-called ELSI questions (ethical, legal, and social implications). In the context of the EMOIO project, ELSI questions are being closely examined through accompanying research.

Sources and literature

[1] R. Parasuraman and M. Rizzo, Eds., Neuroergonomics: the brain at work. New York: Oxford University Press, 2008.
[2] M. Hassenzahl, „User experience (UX): towards an experiential perspective on product quality," 2008, p. 11.

---

[8] http://newsroom.jaguarlandrover.com/en-in/jlr-corp/news/2015/06/jlr_road_safety_research_brain_wave_monitoring_170615/

[9] http://cnbi.epfl.ch/page-81043-en.html

[10] https://www.emotiv.com/

[3] K. Pollmann, M. Vukelić, N. Birbaumer, M. Peissner, W. Bauer, and S. Kim, „fNIRS as a Method to Capture the Emotional User Experience: A Feasibility Study," in Human-Computer Interaction. Novel User Experiences, vol. 9733, M. Kurosu, Ed. Cham: Springer International Publishing, 2016, pp. 37–47.

[4] S. K. Kane, J. O. Wobbrock, and I. E. Smith, „Getting off the treadmill: evaluating walking user interfaces for mobile devices in public spaces," 2008, p. 109.

[5] G. Lehmann, M. Blumendorf, and S. Albayrak, „Development of context-adaptive applications on the basis of runtime user interface models," 2010, p. 309.

[6] J. Nichols and B. A. Myers, „Creating a lightweight user interface description language: An overview and analysis of the personal universal controller project," ACM Trans. Comput.-Hum. Interact., vol. 16, no. 4, pp. 1–37, Nov. 2009.

[7] N. Birbaumer et al., „A spelling device for the paralysed," Nature, vol. 398, no. 6725, pp. 297–298, Mar. 1999.

[8] A. Ramos-Murguialday et al., „Brain-machine interface in chronic stroke rehabilitation: A controlled study: BMI in Chronic Stroke," Ann. Neurol., vol. 74, no. 1, pp. 100–108, Jul. 2013.

[9] A. Kübler et al., „Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface," Neurology, vol. 64, no. 10, pp. 1775–1777, May 2005.

[10] J. I. Münßinger et al., „Brain Painting: First Evaluation of a New Brain–Computer Interface Application with ALS-Patients and Healthy Volunteers," Front. Neurosci., vol. 4, 2010.

[11] M. Bensch et al., „Nessi: An EEG-Controlled Web Browser for Severely Paralyzed Patients," Comput. Intell. Neurosci., vol. 2007, pp. 1–5, 2007.

[12] N. Birbaumer, S. Ruiz, and R. Sitaram, „Learned regulation of brain metabolism," Trends Cogn. Sci., vol. 17, no. 6, pp. 295–302, Jun. 2013.

[13] S. Ruiz et al., „Acquired self-control of insula cortex modulates emotion recognition and brain network connectivity in schizophrenia," Hum. Brain Mapp., vol. 34, no. 1, pp. 200–212, Jan. 2013.

[14] S. W. Choi, S. E. Chi, S. Y. Chung, J. W. Kim, C. Y. Ahn, and H. T. Kim, „Is alpha wave neurofeedback effective with randomized clinical trials in depression? A pilot study," Neuropsychobiology, vol. 63, no. 1, pp. 43–51, 2011.

[15] T. O. Zander and C. Kothe, „Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general," J. Neural Eng., vol. 8, no. 2, p. 025005, Apr. 2011.

[16] C. Dijksterhuis, D. de Waard, K. A. Brookhuis, B. L. J. M. Mulder, and R. de Jong, „Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns," Front. Neurosci., vol. 7, 2013.

[17] C. Berka et al., „EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," Aviat. Space Environ. Med., vol. 78, no. 5 Suppl, pp. B231-244, May 2007.

[18] S. Haufe et al., „Electrophysiology-based detection of emergency braking intention in real-world driving," J. Neural Eng., vol. 11, no. 5, p. 056011, Oct. 2014.

[19] T. O. Zander, L. R. Krol, N. P. Birbaumer, and K. Gramann, „Neuroadaptive technolo- gy enables implicit cursor control based on medial prefrontal cortex activity," Proc. Natl. Acad. Sci., p. 201605155, Dec. 2016.

[20] R. W. Picard, E. Vyzas, and J. Healey, „Toward machine emotional intelligence: analysis of affective physiological state,“ IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[21] H. Förstl, Ed., Theory of mind: Neurobiologie und Psychologie sozialen Verhaltens, 2., Überarb. und aktualisierte Aufl. Berlin: Springer, 2012.

[22] T. O. Zander, J. Brönstrup, R. Lorenz, and L. R. Krol, „Towards BCI-Based Implicit Control in Human–Computer Interaction,“ in Advances in Physiological Computing, S. H. Fairclough and K. Gilleade, Eds. London: Springer London, 2014, pp. 67–90.

[23] K. Pollmann, D. Ziegler, M. Peissner, and M. Vukelić, „A New Experimental Paradigm for Affective Research in Neuro-adaptive Technologies,“ 2017, pp. 1–8.

[24] K. Pollmann, M. Vukelic, and M. Peissner, „Towards affect detection during human-technology interaction: An empirical study using a combined EEG and fNIRS approach,“ 2015, pp. 726–732.

[25] M. Mather and L. L. Carstensen, „Aging and motivated cognition: the positivity effect in attention and memory,“ Trends Cogn. Sci., vol. 9, no. 10, pp. 496–502, Oct. 2005.

[26] M. Vukelić, K. Pollmann, M. Peissner, „Towards brain-based interaction between humans and technology: does age matter?,“ 1st International Neuroergonomics Conference, Oct. 2016.

[27] C. Brunner et al., „BNCI Horizon 2020: towards a roadmap for the BCI community,“ Brain-Comput. Interfaces, vol. 2, no. 1, pp. 1–10, Jan. 2015.

[28] A. Sonnleitner, et al., „EEG alpha spindles and prolonged brake reaction times during auditory distraction in an on-road driving study“. Accid. Anal. Prev. 62, 110–118, 2014

[29] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, „Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear,“ Sci. Rep., vol. 5, no. 1, Dec. 2015

10 Fraunhofer Additive Manufacturing Alliance

From data straight to highly complex products

Dr. Bernhard Müller
Fraunhofer Institute for Machine Tools and Forming Technology IWU
Spokesperson for the Fraunhofer Additive Manufacturing Alliance

Summary

Additive manufacturing is known as 3D printing in popular science. It refers to a relatively new group of manufacturing techniques with unique properties and possibilities compared with conventional manufacturing technologies. The Fraunhofer Additive Manufacturing Alliance currently coordinates 17 Fraunhofer institutes working on additive manufacturing. It covers the entire process chain: the development, application and implementation of additive manufacturing techniques and processes as well as the relevant materials. This chapter provides an overview of the technologies, applications, particular opportunities and further goals of applied research in the area of additive manufacturing within the Fraunhofer-Gesellschaft. We make particular mention of mesoscopic lightweight design, biomimetic structures, high-performance tools for hot sheet metal forming, ceramic components, printable biomaterial, large-size plastic components, integrating sensory-diagnostic and actuator therapeutic functions into implants, and three-dimensional multimaterial components.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_10

10.1 Introduction: history of additive manufacturing

Additive manufacturing (AM), often referred to as 3D printing in popular science, is a comparatively new group of manufacturing techniques with unique properties and possibilities compared to conventional manufacturing technologies we know today. During the early days of additive manufacturing in the 1980s, mainly polymers were being processed. Today, however, metals and ceramic are also being used. Until now, the technology for all of the additive manufacturing techniques has been based on a *layer-by-layer build-up of components*. Originally, additive manufacturing techniques were used for quickly producing prototypes and were referred to as such ("rapid prototyping"). Now, however, further development has made the direct manufacturing of serial components and end products ("direct digital manufacturing") possible.

Additive manufacturing techniques are primarily employed for three reasons:

- Individual item and short-run batch production can often be more economically attractive when molds and tools are avoided.
- Fewer manufacturing restrictions (accessibility for tools, demolding ability, etc.) mean that delicate and highly structured components can be produced, e.g. with anisotropic, locally-varying or functionally integrative properties and movable components.
- Personalized solutions (customization) can be implemented where products are tailored to user or application-specific requirements (e.g. prostheses, shoes).

The last two points are the key drivers today, contributing to the increasing spread of additive manufacturing as an alternative production technique. The central challenge here is to master the competition regarding cost and quality of established batch production processes such as machining and injection molding, and to significantly increase process efficiency (energy use, waste generation, and robustness). This is particularly the case where there are high demands in terms of surface quality and component failure on the application side, such as in aerospace and mechanical engineering, and also in the case of large volumes of personalized mass-produced products (mass customization, e.g. of glasses or shoes).

Since 2005 development, and since 2009 market activity have been strongly influenced by two trends: the increasing activities of open source communities (in particular the RepRap project), and the Fab@Home concept (desktop printing such as MakerBot). The fascination with additive manufacturing techniques, the desire to participate in production processes, the opportunity to produce replacement parts on demand, and the reintegration of consumer product manufacturing into local economies are all important drivers of this development. Whereas this development prin-

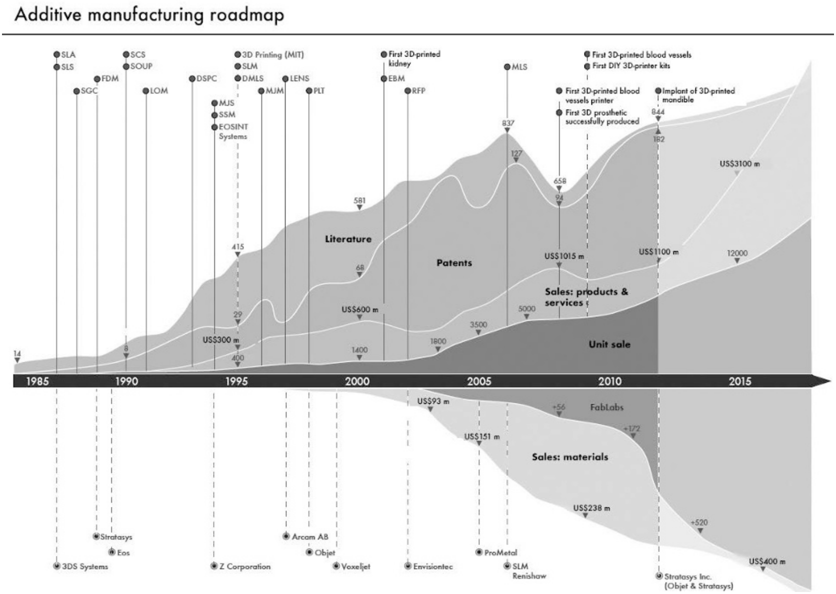


Fig. 10.1 Graphical roadmap of additive manufacturing to the present (Fraunhofer UMSICHT, Fraunhofer Additive Manufacturing Alliance 2012)

cipally addresses additive manufacturing techniques for polymers, metal techniques remain reserved for industrial applications so far. Here, however, unusual innovative dynamism and extraordinary growth rates are to be noted, driven by industrial application sectors such as aerospace, energy engineering, medical technology and tooling/mold and die making. The appeal of additive manufacturing has fundamentally grown over the last three decades. This can be seen from numerous indicators including the frequency of patents and publications, revenue from machines and materials, and the founding of companies and new informal communities (see Fig. 10.1).

The current focus within industrial applications, however, remains largely restricted to metal and plastic materials.

10.2 Additive manufacturing at Fraunhofer

The roots of the Fraunhofer Additive Manufacturing Alliance stretch back to the year 1998 when the Rapid Prototyping Alliance was born. Formally relaunched as the Fraunhofer Additive Manufacturing Alliance in 2008 with eight member insti-

Fig. 10.2 Members of the Fraunhofer Additive Manufacturing Alliance (Fraunhofer Additive Manufacturing Alliance 2017)

tutes, it now comprises 17 (cf. Fig. 10.2) and thus reflects the dynamic global development of this manufacturing technique.

The Fraunhofer Additive Manufacturing Alliance’s comparative global position is measured against various criteria and was assessed as part of a competitor analysis carried out by the Fraunhofer ISI, INT, IAO, and IMW institutes on behalf of Fraunhofer head office [1]. To do this, researchers used first and foremost certain specific indicators to review the appeal of the research field and Fraunhofer’s relative position in comparison to other research bodies. The researchers then used future-oriented studies and publications to assess the long-term potential for applications and research and estimated future market potential and research market dynamics by analyzing patents and publications. They calculated Fraunhofer’s current position compared with other research bodies based in part on Fraunhofer’s patent and publication activities compared with other research institutes. In addition, they carried out a brief survey at the Fraunhofer institutes, which are members of the Additive Manufacturing Alliance or which have been active in publishing in this field. The breadth of the additive manufacturing technologies researched and materials produced here can be seen in Tables 10.1 and 10.2.

Table 10.1 Additive manufacturing techniques in use at Fraunhofer [1]

	Technique/ Institute	VP	MJ	PBF	SL	ME	BJ	DED	Other
Fraunhofer Additive Manufacturing Alliance	IFAM	x		x		x	x		x
	IKTS	x	x	x		x	x		
	IFF					x	x		x
	IPT							x	x
	IPA	x	x	x		x	x		x
	ILT	x		x				x	
	IWM	(x)^{1}	(x)^{1}	(x)^{1}	(x)^{1}	(x)^{1}	(x)^{1}		(x)^{1}
	IWU			x		x			
	UMSICHT			x		x			
	IGD	(x)^{2}	(x)^{2}	(x)^{2}	(x)^{2}	(x)^{2}	(x)^{2}		(x)^{2}
	IGB	x	x			x			x
	EMI			x		x			
	IST								
	IGCV			x		x	x		
	IAO								
	IWS	x		x	x	x		x	x
	IPK			x				x	
(Previous) publications in AM field but not part of Alliance	ISC	x					x		x
	ICT					x			
	IOF								
	IAP								

(x) R&D contributions only

(x)^{1} Working on the mechanical and tribological characterization of additively manufactured components, the design of components for additive manufacturing and the simulation of process steps.

(x)^{2} Development of algorithms and software for controlling 3D printers (no materials development, but optimization of materials and component properties through adaptation of process parameters)

Legend:

VP – Vat Photopolymerization: selective light curing of a liquid photopolymer in a vat, e.g. stereolithography (SLA/SL)

MJ – Material Jetting: drop-by-drop application of liquid material, e.g. multijet modeling, polyjet modeling

PBF – Powder Bed Fusion: selective melting of regions within a powder bed, e.g. laser sintering (LS), beam-based melting (LBM, EMB), selective mask sintering

SL – Sheet Lamination: successive layering and bonding of thin sheets of material, e.g. layer laminated manufacturing (LLM), laminated object manufacturing (LOM) also: stereolithography (SLA)

ME – Material Extrusion: targeted deposition of material through a nozzle, e.g. fused layer modeling (FLM), fused deposition modeling (FDM)

BJ – Binder Jetting: selective adhesion of powdery material using a liquid binder, e.g. 3D printing (3DP)

DED – Directed Energy Deposition: targeted welding of the material during deposition, e.g. laser powder build-up welding (Laser-Pulver-Auftragschweißen – LPA), direct metal deposition (DMD), laser cladding

Table 10.2 Additive materials in use at Fraunhofer [1]

	Materials/ Institute	Plastics	Metals/ alloys	Ceramics	Composites	Biol. mat.	Other
Fraunhofer Additive Manufacturing Alliance	IFAM		D/A	D	D		
	IKTS		D/A	D/A	D/A		
	IFF	A				A	
	IPT		D/A		A		
	IPA	D/A		D/A	D/A	D/A	D/A
	ILT	D	A	D			
	IWM						
	IWU	D/A	D/A		D	D	
	UMSICHT	D/A			D		
	IGD	D/A[2]	D/A[2]		D/A[2]		
	IGB	D/A			D/A	D/A	D/A Functional nanoparticles (metal oxides)
	EMI	A	D/A		A		
	IST	D/A	D/A		D/A	D/A	Combination process (plastics printing and plasma treatment)
	IGCV	D/A	D/A	D/A	D/A		
	IAO						
	IWS	D/A	D/A	D/A	D/A		
	IPK		D/A	D/A			
(Previous) publications in AM field but not part of Alliance	ISC		A	D/A			
	ICT	D/A		D/A			
	IOF						
	IAP						

Legend
D = Development; A = Application
[2] Development of algorithms and software for controlling 3D printers (no materials development, but optimization of materials and component properties through adaptation of process parameters)

This integrated view and assessment takes account both of the appeal of the technological field of additive manufacturing/the individual technological subthemes as well as of the positioning of Fraunhofer within this field. The following criteria were used to identify Fraunhofer's position (largely defined by the Alliance's member institutes):

- Publically funded projects (Nationally: ranked first by a large margin according to number of projects and total amount; EU: ranked second by total amount, first by number of projects)
- Patent activity (ranked 21st across all additive manufacturing technologies according to analysis of patent family registration between 2009 and 2014; ranked first globally among research institutes; across all technologies among the top 10 research institutes)
- Publication activity (ranked first in Germany and fourth for global scientific publications in peer-reviewed journals; ranked between first and sixth for conference papers, non-peer-reviewed publications and press releases)
- Networking in the scientific community (close networks with players with institutional connections, e.g. professorships of heads of institutes; diverse network with European players in particular, but also with American and select Chinese players but a lack of clear and well-developed network with many players)

On the basis of these assessments, it can be concluded that Fraunhofer is the world's most broadly-positioned research player in the field of additive manufacturing. Its network clearly concentrates on industry companies.

Alongside Fraunhofer's unique position of being active in all of the technology fields, for certain technologies (powder bed fusion, material jetting, and binder jetting) Fraunhofer is among the leading players [1].

The scientific excellence of the Additive Manufacturing Alliance is reflected, along with the aforementioned aspects, in the specialist international Fraunhofer Direct Digital Manufacturing Conference DDMC, organized by the Alliance since 2012, every two years in March, in Berlin. The research findings of the Alliance institutes presented here, the renowned global keynote and specialist speakers, and the large number of conference delegates show Fraunhofer's outstanding worldwide reputation in the field of additive manufacturing.

The Alliance strives to play a leading global role in applied additive manufacturing research. Its focus, then, is on combining the strengths of the Alliance's members and using the various complementary skills to provide an appealing offer of comprehensive commissioned research to industrial customers. The Alliance's research spectrum here stretches right across the entire field of additive manufacturing, in a very comprehensive way, and can essentially be divided into four key areas or research focus areas:

- Engineering (application development)
- Materials (plastics, metals, ceramics)
- Technologies (powder bed, extrusion and printing based)
- Quality (reproducibility, reliability, quality management)

The following selected example projects provide an insight into the diversity of applied research in additive manufacturing (3D printing) at Fraunhofer.

10.3 Additive manufacturing – the revolution of product manufacturing in the digital age

Prof. Dr. Christoph Leyens · Prof. Dr. Frank Brückner · Dr. Elena López ·
Anne Gärtner
Fraunhofer Institute for Material and Beam Technology IWS

The AGENT-3D joint research project coordinated by the Fraunhofer Institute for Material and Beam Technology IWS aims to develop solutions to existing scientific and technical, political and legal, and socioeconomic challenges for additive manufacturing together with more than 100 project partners, mostly coming from industry.[1]

Following the completion of the strategy phase and the construction of a measuring and testing center with modern equipment for, among other things, optical and x-ray examination and measurement of additively manufactured components (e.g. using scanners or computer tomography), the consortium is working nowadays on implementing the strategical roadmap by means of basic projects and over 15 technology projects (further technology projects will complete the strategical roadmap until the year 2022). These projectsfocus, for example, on integrating functionalities into components, combining conventional and additive manufacturing, allowing the processing of multimaterials and enhancing the material portfolio for additive processes, and last but not least, quality management along the whole process chain.

Additive manufacturing allows complex components to be constructed layer-by-layer directly based on digital data (cf. Fig. 10.3). The principle of reverse engineering, on the other hand, allows a scaled-up reproduction of an original part

---

[1] AGENT-3D, BMBF, *Zwanzig20 – Partnerschaft für Innovation* ("Twenty20 – partnership for innovation") program (BMBF-FKZ 03ZZ0204A)

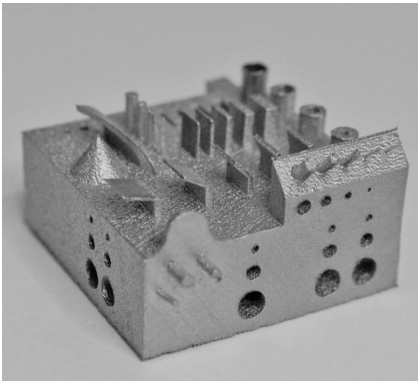


Fig. 10.3 Demonstrator from the AGENT_3D_Basis subproject, manufactured using laser beam melting, to illustrate challenging shapes (Fraunhofer IWS)

Fig. 10.4 Reproduction following scanning of a bird skull in original size and at ten times actual size (Fraunhofer IWS)

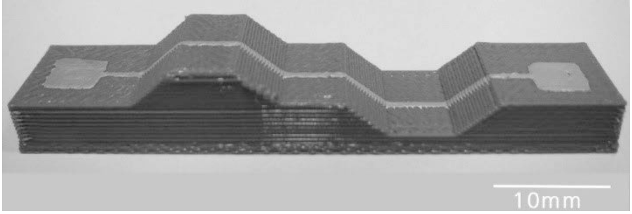


Fig. 10.5 Printed conductor track from the AGENT-3D_elF technology project (Fraunhofer IWS)

Also redesign of the part to fulfill special requirements can be easily done in this way. Scans are used to provide data from which identical or optimized parts in terms of i.e. design can be printed (cf. Fig. 10.4). In addition, a function integration can be included, for instance in terms of electronic functionalities such as conductor tracks (cf. Fig. 10.5) or sensors that can be directly printed into three-dimensional components. In this way, digitization enables completely new possibilities for designing and manufacturing products.

Contact:
Dr. Elena López, elena.lopez@iws.fraunhofer.de, +49 351 83391-3296
www.agent3D.de

10.4 Mesoscopic lightweight construction using additively manufactured six-sided honeycombs

Matthias Schmitt · Dr. Christian Seidel
Fraunhofer Research Institution for Casting, Composite and Processing Technology IGCV

Lightweight construction plays an important role in the aerospace and automotive industries in order to reduce the energy demand during operation or to raise the performance of the overall system. Beyond this, principles of lightweight construction are utilized in all sectors of industry to achieve ecological and economic use of raw materials. Nevertheless, ideal lightweight construction designs can often not be implemented because the corresponding manufacturing technologies for materialization are not available. Additive manufacturing techniques such as laser beam melting (LBM) can provide assistance here. The process operation can permit the manufacturing of geometrically complex components in small batches, highly efficiently.

Fraunhofer IGCV is working on optimizing lattice and honeycomb structures for sandwich components. In sandwich components, a lightweight core is supplied with solid, rigid covering layers producing a material compound that demonstrates significantly better mechanical properties than the sum of the individual layers. Honeycomb structures are thus particularly well suited for use as core material for high-strength lightweight constructions since the hexagonal geometry allows maximum compression loads to be absorbed with minimal core weight. Using conventional methods to manufacture honeycomb structures, however, produces significant limitations with respect to fully exploiting the potential of lightweight construction. This is due, on the one hand, to the fact that conventional, e.g. forming, manufacturing techniques produce regular material filling degrees that do not allow for load optimization of the structure. Conventionally manufactured honeycomb structures, for example, thus offer almost no possibility for placing more material at points of high load and reducing the material thickness of the honeycomb walls at points of low load. In addition to this, conventional techniques offer limited suitability for adapting the honeycomb structures to freeform surfaces. By using additive manufacturing on the other hand, honeycomb structures can be adapted to complex geometries (cf. Fig. 10.6).

To achieve this, Fraunhofer IGCV developed a software tool for the CAD program Siemens NX, which aligns the honeycomb with a given freeform surface and dimensions the honeycomb's individual segments in keeping with the load. In addition, inserts can be provided to introduce threads into the sandwich composite, for

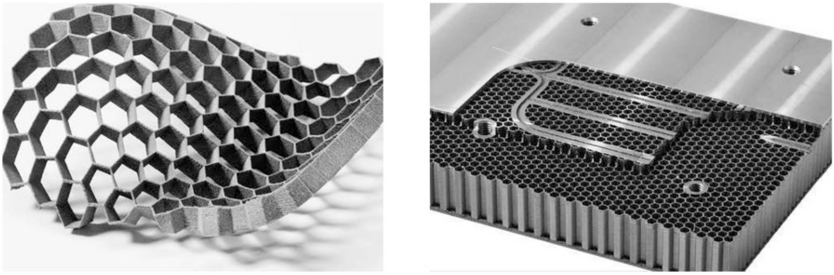


Fig. 10.6 Honeycomb structure adapted to a freeform surface, and honeycomb with load application elements (Fraunhofer IGCV)

example (cf. Fig. 10.6). By using powder bed-based additive manufacturing processes, the honeycomb structures produced in this way were able to be generated in both plastic as well as metal. Further potential for reducing weight and increasing rigidity lies in the combination of additively manufactured honeycomb structures with a covering layer of carbon fiber-reinforced plastic.

Contact:
Dr. Christian Seidel, christian.seidel@igcv.fraunhofer.de, +49 821 90678-127

10.5 Using biomimetic structures for esthetic consumer goods

Dr. Tobias Ziegler
Fraunhofer Institute for Mechanics of Materials IWM

Together with our project partners Industrial Design, Folkwang University Essen, Fraunhofer UMSICHT, Sintermask GmbH, rapid.product manufacturing GmbH and Authentics GmbH, a numeric tool for design, analysis and optimization was developed at Fraunhofer IWM.[2] The tool fills specified external shapes with a cellular structure based on trabecular cells, similar to cancellous bone. For additive manufacturing to be cost-effective, it is important to be able to assess the mechanical properties of products without having to produce additional exemplars for mechan-

---

[2] Bionic Manufacturing, DLR, part of the BMBF's Biona program (BMBF-FKZ 01RB0906)

Fig. 10.7 *Cellular Loop*, a mechanically developed and manufactured designer cantilever chair. Photo by Natalie Richter (Folkwang University of the Arts)

ical testing. Due to the regularity of the cellular structure, this approach facilitates the advance calculation of mechanical properties such as load-bearing capacity and rigidity. Experimental data from just a few representative samples is the only input parameter needed to characterize the material and the process for finite element models. Any additive manufacturing technique and material can be used here.

In order to optimize the component's mechanical properties, the trabecular cell's microstructure can be adapted to a given load. This is done by locally anisotropically increasing the diameter of the trabecular rods. In this way, the load-bearing capacity of the component can be significantly raised, with minimum material use and production time.

The tool presented can be used on a large number of components and allows mechanical properties to be calculated and improved. Due to its visual properties, the biomimetic cellular structure also leads to esthetically pleasing products, as shown by the product described in what follows.

As a demonstrator, a bionic cantilever chair was developed by the group around Anke Bernotat at the Folkwang University of the Arts. The loads produced by an individual sitting on it were calculated at Fraunhofer IWM. Next, the microstructure was adapted to this load. The shape was divided into producible segments and the chair was then manufactured in selective-laser-sintering by our partners at rpm-factories. The chair provided the expected load-bearing capacity and also corresponds to the highest esthetic standards.

Contact:
Dr. Tobias Ziegler, tobias.ziegler@iwm.fraunhofer.de, +49 761 5142-367

10.6 High-performance tools for sheet metal hot forming using laser beam melting

Mathias Gebauer
Fraunhofer Institute for Machine Tools and Forming Technology IWU

Manufacturing complex sheet metal parts from high-strength steel places large demands on cold forming. High pressing forces and the relatively high springback represent huge challenges. Very rigid tools made of expensive material are required that are nevertheless subject to increased wear. An alternative to cold forming is sheet metal hot forming or press hardening. Here, the sheet metal blank is heated above the austenitizing temperature (above 950 °C) and rapidly cooled to below 200 ºC during forming, creating a martensitic structure.

The structure of a hot forming tool is more complex than that of a conventional one. The reason for this is the necessary integration of cooling channels into punch and die. The channels, generally produced by deep drilling, are limited in their minimally representable diameters, which has a direct impact on the contour dis-

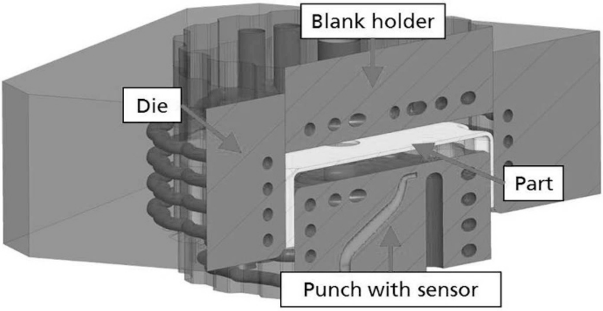


Fig. 10.8 3D CAD model of a press hardening tool with conformal cooling channels (Fraunhofer IWU)

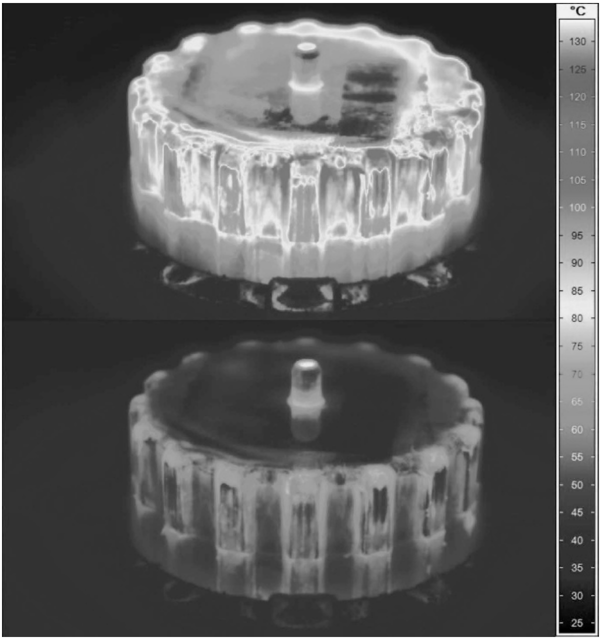


Fig. 10.9 Thermographic image of a tool punch (Fraunhofer IWU)

tance achievable. For this reason, targeted tempering of individual regions conformal to the contour of the tool is only achievable with great effort and significant limitations for hot forming tools. This often causes insufficient target temperature achievement and too little heat dissipation in the tool's critical regions.

As part of the additively manufactured HiperFormTool[3] project, research has been carried out into how sheet metal hot forming can be made more efficient by means of additively manufactured active tool components. To achieve this, the thermal behavior of the tools and of the forming process was precisely analyzed via simulation and various cooling channel geometries compared. Based on the results of the simulation and by using the geometric freedom of additive laser beam melting, an innovative and contour-close cooling system was developed. The primary goal of the research studies was to significantly shorten the cycle time. In addition, a concept for sensor integration during the additive manufacturing process was developed and implemented.

---

[3] HiperFormTool, high-performance sheet metal hot forming tools using laser beam melting, ERA-NET joint project, MANUNET HiperFormTool (BMBF-FKZ 02PN2000)

The innovative cooling system allowed a significant reduction of holding time in press hardening of 70%, from 10 s to 3s, with hot-formed components of identical precision and hardness. In total, more than 1,500 components were formed and 3 hours of manufacturing time saved in the process. The function of the firmly bonded integrated thermosensors could be proven through the precisely documented temperature progress during the laser beam melting process itself, the heat treatment, and the actual forming tests.

Contact:
Mathias Gebauer, mathias.gebauer@iwu.fraunhofer.de, +49 351 4772-2151

10.7 Additive manufacturing of ceramic components

Uwe Scheithauer · Dr. Hans-Jürgen Richter
Fraunhofer Institute for Ceramic Technologies and Systems IKTS

In contrast to the additive manufacturing of polymer or metal components, in ceramic component manufacture the typical heat treatment processes such as debinding and sintering follow the actual additive manufacturing process (shaping). Here, the organic additives are first removed from the additively manufactured green body

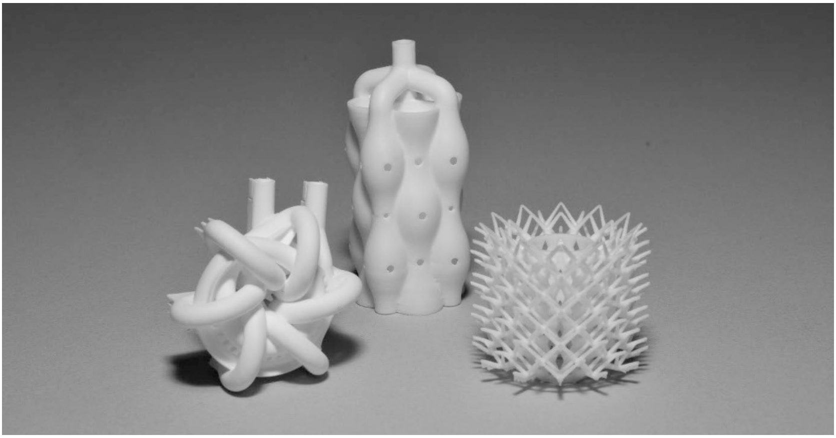


Fig. 10.10 Additively manufactured aluminum oxide components for applications as heat exchangers or mixers for two fluids (Fraunhofer IKTS)

before the ceramic particles are sintered, which generally involves a significant reduction in volume, at temperatures above 1000 °C. It is only during the sintering phase that the component gains its final properties.

For manufacturing ceramic components additively, different processes are used that can generally be divided into powder bed- and suspension-based or indirect processes (areal application of the material and selective solidification) and direct processes (selective application of the material).

With suspension-based processes, the base materials are in the form of suspensions, pastes, inks, or semi-finished products such as thermoplastic feedstocks, green sheets, or filaments. Compared with powder bed-based processes, higher green densities are achieved with suspension-based additive manufacturing processes, which then lead to a dense microstructure in the sintered part and decreased surface roughness. New, complex ceramic structures illustrate the potential of additive manufacturing for ceramic (cf. Fig. 10.10).

One current focus is on the development of processes for manufacturing multi-material compounds (e.g. ceramic/metal) and components with a gradient of properties (e.g. porous/dense). The direct processes in particular offer huge potential here due to the selective application of different materials. In this way, in future it will be possible to manufacture components with highly complex inner and outer geometries that will also combine the properties of various materials (e.g. electrically conductive/non-conductive, magnetic/non-magnetic).

Materials, equipment, process, and component development issues are being worked on together with national and international partners in several BMBF pro-

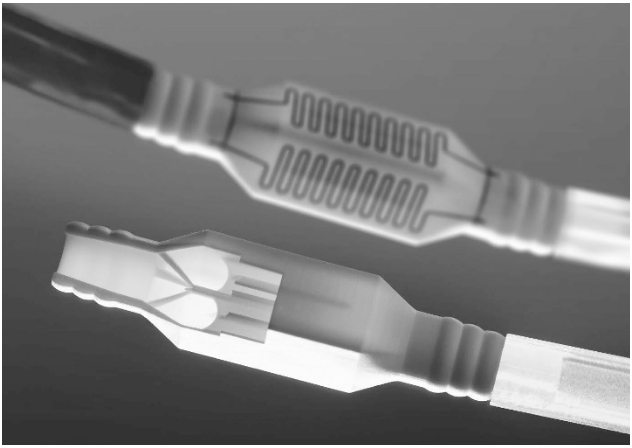


Fig. 10.11 Ceramic heating element structure, additively manufactured and functionalized using aerosol printing (Fraunhofer IKTS)

jects (AGENT-3D: IMProve + MultiBeAM + FunGeoS; AddiZwerk) and within the EU project cerAMfacturing. An additional area of focus is the development of hybrid processes, where additive and conventional manufacturing techniques are combined. Using this technique, it is possible to further customize components that are mass produced, or to further functionalize additively manufactured components (cf. Fig. 10.11). Alongside the development and adaptation of the process, the constant broadening of the useable material portfolio is obviously also an indispensable development task.

Contact:
Dr. Hans-Jürgen Richter, Hans-Juergen.Richter@ikts.fraunhofer.de,
+49 351 2553-7557

10.8 Printable biomaterials

Dr. Kirsten Borchers · Dr. Achim Weber
Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB

Printing biological and biofunctional materials – also known as bioprinting – is a relatively new and promising option for giving surfaces a function or manufacturing entire 3D objects (cf. Fig. 10.12). Current research and development studies that Fraunhofer IGB is involved in provide fuel to the vision of one day using customized biological implants.

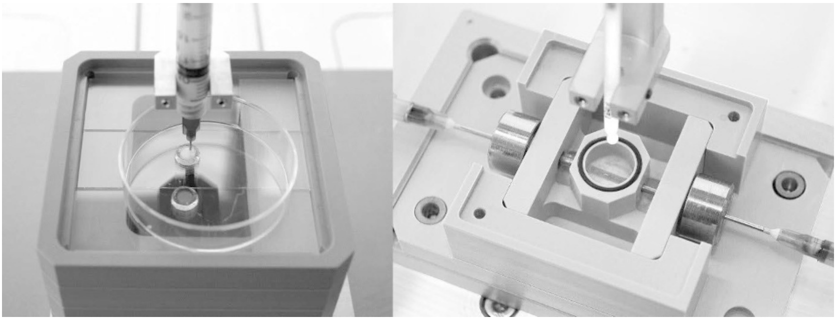


Fig. 10.12 Bioinks made of modified biomolecules are designed for digital generation of biological tissue replacements. The modified biomolecules at Fraunhofer IGB can be formulated to form low or high viscous fluids and composed to host different cell types. Left: viscous ink for bone cells, right: soft ink for sensitive fat cells.

Various printing techniques such as inkjet or dispensing processes require different rheological material properties. At the same time, the so-called bio-inks must be stabilized after printing so that the desired biological functions are available.

Biopolymers are optimized by nature and fulfill complex tasks: as matrices of tissue they harbor living cells for example; they store water and water-soluble substances and release them on demand; and they are involved in the transmission of biological signals. These extensive functions cannot simply be reproduced via chemical synthesis, but it is possible to chemically modify suitable biomolecules and thus make them usable for digital printing processes.

Fraunhofer IGB uses biopolymers from the extracellular matrix of natural tissues such as gelatin as a derived product from collagen, heparin, hyaluronic acid, and chondroitin sulfate, and provides them with additional functions. By “masking” specific functional groups, for example, intermolecular interactions can be reduced and the viscosity and gelatinization behavior of the biopolymer solutions thus influenced. In addition, reactive groups can be adapted in order to fix biomolecules onto surfaces and produce hydrogels of variable strength and swelling capacity, see Fig. 10.13. [2][3][4][5] Finally, by means of the formulation – that is, the mixing and addition of signal substances or biofunctional particles – printable biomaterials with tailored properties are produced. [6][7]

With chemically modified biopolymers as a basis, Fraunhofer IGB develops printable biomolecule solutions, bio-based release systems, and cell-specific matrices for tissue regeneration. [8][9][10]

Contact:
Dr. Achim Weber, achim.weber@igb.fraunhofer.de, +49 711 970-4022

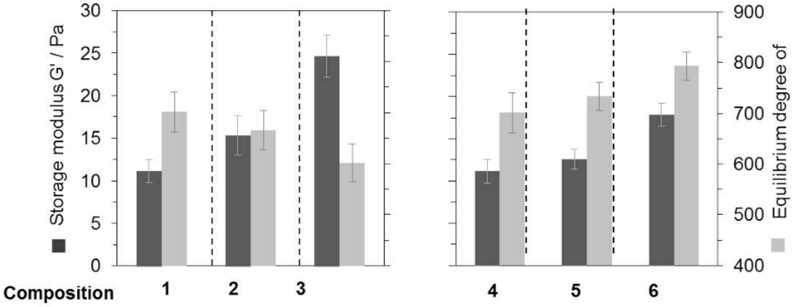


Fig. 10.13 Via the formulation of differently modified biomolecules, hydrogels with the same biopolymer concentration and composition can be produced with different combinations of properties (Fraunhofer IGB).

10.9 Development and construction of a highly productive manufacturing facility for additive manufacturing of large-scale components made of arbitrary plastics

Dr. Uwe Klaeger
Fraunhofer Institute for Factory Operation and Automation IFF

In many sectors, manufacturing large components is associated with high production costs. In order to be able to produce these kinds of elements economically, high build-up rates (production speeds) and simultaneously low material costs are required.

One promising approach to solving this problem is the development of a new, inexpensive plant concept, which is being pursued within the High Performance 3D joint project[4] by six industrial companies and three research institutes.

The technology combines additive manufacturing techniques with modern industrial robotics, thus facilitating the economical manufacture of individual components of arbitrary sizes and weights.

The fundamental idea behind this process is based on the combination of a special granulate extruder with a flexible buckling arm robot. The highly productive plant uses three extruders that can be utilized to apply different materials layer-by-layer. The materials palette includes both hard/soft combinations and different colors as well as materials filled with glass or carbon fibers. The extruder unit was designed for a maximum component build-up rate of 2 kg/h for standardized granulates, typical plastic materials such as ABS, PMMA, PP, PC, PC/ABS, and PLA. To guarantee a continual flow of material, a modified needle nozzle was fitted to prevent uncontrolled filament formation during the construction process. By constantly measuring the online temperature in the installation room, stable viscosity behavior of the plastics is achieved.

Component construction takes place on a heated work platform mounted to the robot. In order to produce a three-dimensional part without anisotropy, the construction platform moves on six axes so that the material application point is always perpendicular to the extruder nozzle. A second robot places additional elements into the component (including metallic ones), thus facilitating the automatic integration of additional functional elements. The prototype plant is initially designed for com-

---

[4] HP3D: Concept development and construction of a highly productive manufacturing plant for additive manufacturing of large-scale components made of arbitrary plastics-High Performance 3D (BMBF-FKZ 02P14A027)

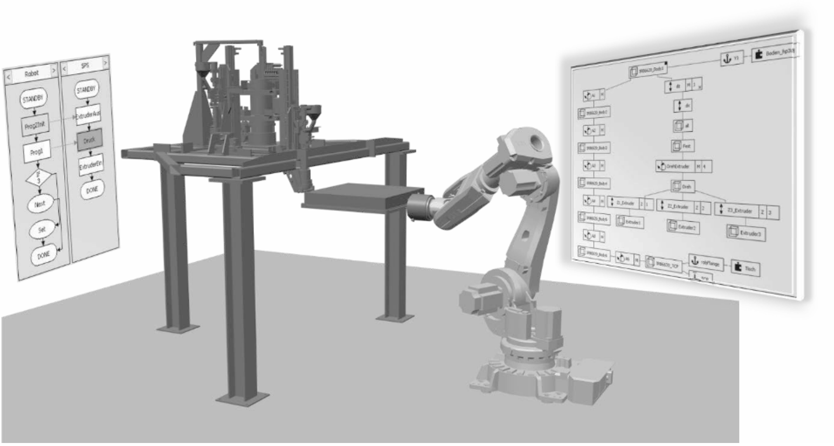


Fig. 10.14 Additive manufacturing of large components using universal industrial robotics and simulation tools to assist development (Fraunhofer IFF)

ponent volumes of 1000 x 1000 x 1000 mm³ and maximum component weights of 25 kg.

A key element of the overall technological concept is the simulation of complex production processes associated with development. To do this, the VINCENT universal simulation tool developed by Fraunhofer IFF is used (cf. Fig. 10.14). The results of the simulation are directly incorporated into the constructive development of the manufacturing plant. The program facilitates process visualization and reachability testing of all paths for the layer-by-layer component construction as well as a collision detection in the workspace. In this way, a geometric and functional test of the plant is possible even before commencing the manufacture of its components so that a significant reduction in the plant development and commissioning time is achieved.

The new 3D printing process combining extruders and robots opens up new possibilities in the production of large, complex, plastic parts. Since expensive molds or tools are not required, component production is subject to hardly any spatial limitations. With the significantly shorter process chain, large components will in future be able to be produced economically and flexibly, leading to a variety of new or improved products in a large range of market segments.

Contact:
Dr. Uwe Klaeger, uwe.klaeger@iff.fraunhofer.de, +49 391 40 90-809

10.10 Integration of sensory-diagnostic and actuator therapeutic functions in implants

Dr. Holger Lausch
Fraunhofer Institute for Ceramic Technologies and Systems IKTS
Thomas Töppel
Fraunhofer Institute for Machine Tools and Forming Technology IWU

Both diagnostic as well as therapeutic functions are expected from theranostic implants. In engineering terms, then, implants should have sensory and actuator components integrated into them. The advantage of this kind of strategic approach is that treatment-relevant information can be gathered where it arises so that biological treatment effects can be achieved locally precisely there. This strategy was implemented in the Fraunhofer Theranostic Implants lighthouse project in terms of a form-fit, force-fit bonded embedding of actuators and sensors into a compact additively manufactured hip implant. By means of this complete integration in the implant, the measurement of forces or stresses can thus take place directly in the region where they occur within the implant.

For therapeutic functions, the corresponding actuator module, hermetically encapsulated in the interior of the implant, can ensure a partial or total excitation of the implant close to the desired surface area for biomechanical, electrical, or chemical

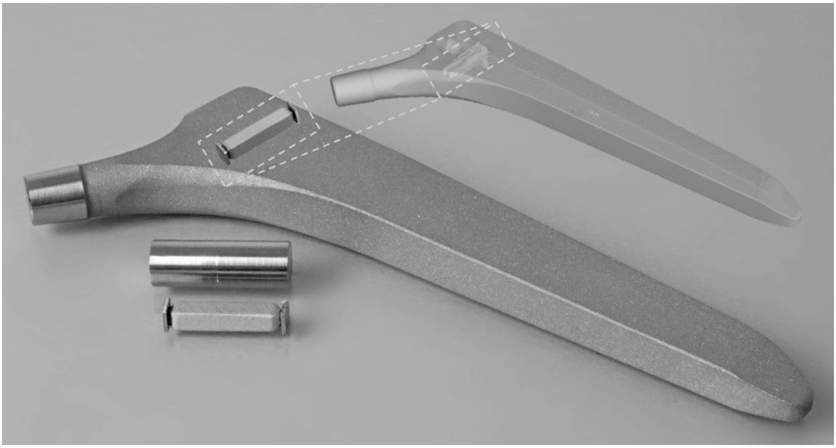


Fig. 10.15 Hip stem implant with integrated sensor-actuator unit; top right: CT image (Fraunhofer IWU)

excitation of the interface between the implant and the tissue. As a result of this project, it was possible to integrate thermally sensitive functional components into a titanium hip stem implant produced additively via laser beam melting. To do this, both the sensor/actuator and inductor for the wireless energy and data transmission were incorporated into an additively manufactured carrier structure that is welded with the main body of the implant later in the process. Therefor, the inherent properties of the additive manufacturing process are used to apply thermal energy into the material in a spatially and temporally highly limited and highly controlled manner using the laser beam. Combined with a suitable laser beam process control and a specially developed additively manufactured ceramic-metal multilayer protective coating system (ceramic metallic covering – CMC) for the sensors/actuators, it was possible to ensure that the functionality of the sensors/actuators was retained in spite of the high melting temperatures of the TiAl6V4 titanium alloy of almost 1700 °C. The process chain developed for the form-fit, force-fit, firmly bonded integration of the sensors/actuators can be transferred to additional applications and used for component-integrated condition monitoring or actuator functionalization for example.

Contact:
Dr. Holger Lausch, holger.lausch@ikts.fraunhofer.de, +49 341 35536-3401
Thomas Töppel, thomas.toeppel@iwu.fraunhofer.de, +49 351 4772-2152

10.11 Generating three-dimensional multi-material parts

M.Sc. Matthias Schmitt · M.Sc. Christine Anstätt · Dr. Christian Seidel
Fraunhofer Research Institution for Casting, Composite and
Processing Technology IGCV

Fraunhofer IGCV's Additive Manufacturing group is currently working primarily on powder bed-based techniques for producing high-performance metal components such as laser beam melting (LBM). Here, a laser beam is used to selectively melt and solidify thin layers of metal powder. At present, the process can be used to produce components made of a single material. Multi-material components are characterized by at least two different materials that are firmly joined to one another. The manufacture of 2D multi-material components, which feature a change of material between subsequent layers, is already possible by means of time-consuming manual changes of material. At present, this is typically not possible for a 3D multi-material component since both materials must be present within a single

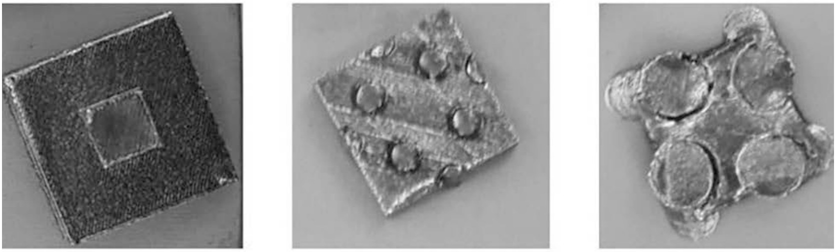


Fig. 10.16 Multi-material structures with 1.2709 tool steel and copper alloy CuCr1Zr (Fraunhofer IGCV)

layer here. To manufacture these parts, it is necessary to adapt the powder application mechanism in order to facilitate the deposition of a second material in the powder layer. For this reason, at Fraunhofer IGCV, a new application mechanism was integrated into an LBM plant via software and hardware so that the construction of 3D multi-material components is now possible in a laser beam melting system.

An initial application of the modified laser beam melting system focused on the production of structures made of 1.2709 tool steel and a copper alloy (cf. Fig. 10.16) within the project ForNextGen supported by the Bavarian Research Foundation. The project consortium consisting of six academic partners and 26 industrial companies has the goal of laying manufacturing science foundations for the use of additive manufacturing processes in mold and tool making, and is being supported by the Bavarian Research Foundation. The classifying and subsequent introduction of these processes is intended to lead to significant improvements in the complexity of shapes, strength, and production time and cost of tools in primary shaping and forming. The multi-material processing researched by Fraunhofer IGCV thus offers great potential for tool shapes and uses. Using the example of a sprue bushing for a die casting mould, a base body of 1.2709 tool steel is constructed and equipped with CuCr1Zr (copper alloy) in two different component regions for improved heat dissipation. By means of these internal cooling structures made of highly thermally-conductive material, the heat balance can be improved and the cycle time thus reduced.

Beyond the work carried out as a part of the ForNextGen project, Fraunhofer IGCV could already show that the current laser beam melting system can even be used to produce multi-material components of metal alloy and a technical ceramic (AlSi12 and Al_2O_3).

Contact:
Christine Anstätt, christine.anstaett@igcv.fraunhofer.de, +49 821 90678-150

Sources and literature

[1] Schirrmeister, E. eta al: Wettbewerbsanalyse für ausgewählte Technologie- und Geschäftsfelder von Fraunhofer – WETTA, Teilbericht: Generative Fertigung. Fraunhofer-Gesellschaft (intern), Januar 2015

[2] EP2621713 B1Vorrichtung und Verfahren zur schichtweisen Herstellung von 3D-Strukturen, sowie deren Verwendung, 2011, FraunhoferGesellschaft.

[3] DE1020112219691B4: Modifizierte Gelatine, Verfahren zu ihrer Herstellung und Verwendung, 2014, Fraunhofer Gesellschaft.

[4] Hoch, E., et al., Chemical tailoring of gelatin to adjust its chemical and physical properties for functional bioprinting. Journal of Materials Chemistry B, 2013. 1(41): p. 5675-5685.

[5] Engelhardt S. et al., Fabrication of 2D protein microstructures and 3D polymer–protein hybrid microstructures by two-photon polymerization, Biofabrication, 2011. 3: p.025003

[6] Borchers, K., et al., Ink Formulation for Inkjet Printing of Streptavidin and Streptavidin Functionalized Nanoparticles. Journal of Dispersion Science and Technology, 2011. 32(12): p. 1759-1764.

[7] Knaupp, M., et al., Ink-jet printing of proteins and functional nanoparticles for automated functionalization of surfaces. 2009.

[8] Wenz, A., et al., Hydroxyapatite-modified gelatin bioinks for bone bioprinting, in BioNanoMaterials. 2016. p. 179.

[9] Huber, B., et al., Methacrylated gelatin and mature adipocytes are promising components for adipose tissue engineering. Journal of Biomaterials Applications, 2016. 30(6): p. 699-710.

[10] Hoch, E., Biopolymer-based hydrogels for cartilage tissue engineering. Bioinspired, Biomimetic and Nanobiomaterials, 2016. 5(2): p. 51-66.

11

Future Work Lab

The workplace of the future

Prof. Dr. Wilhelm Bauer · Dr. Moritz Hämmerle
Fraunhofer Institute for Industrial Engineering IAO
Prof. Dr. Thomas Bauernhansl · Thilo Zimmermann
Fraunhofer Institute for Manufacturing Engineering and Automation IPA

Summary

The Future Work Lab – an innovation laboratory for work, people, and technology – provides companies, associations, coworkers and labor unions with extensive opportunities to experience future-oriented work concepts. The laboratory combines demonstrations of specific Industry 4.0 applications with competency-development offers and integrates the current state of work research. In this way, it facilitates holistic developmental progress in the field of work, people, and technology. Taken as a whole, the Future Work Lab provides a significant contribution to long-term increases in companies' competitiveness through participative design of sustainable working environments..

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_11

Project overview

Partner organizations
Fraunhofer Institute for Industrial Engineering IAO
Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Institute of Human Factors and Technology Management (IAT) at the University of Stuttgart Institute of Industrial Manufacturing and Management (IFF) at the University of Stuttgart

Research schedule, sponsorship
Project duration: 05/2016–04/2019 Sponsorship: €5.64 m.
Sponsor: BMBF
Project manager: PTKA Karlsruhe

Contact
Dr. Moritz Hämmerle
Fraunhofer IAO
Tel. +49 711 970 -2284
moritz.haemmerle@iao.fraunhofer.de
www.futureworklab.de

11.1 Introduction: the digitization and Industry 4.0 megatrend

Our society – and with it, many others in the world – is faced with new and extensive challenges as a result of demographic changes, shortages in specialist staff, and the onward march of digitization. The Internet and digital technologies, first and foremost the mobile use of data and artificial intelligence, are not only reshaping our everyday lives, they are also leading to far-reaching changes in the economy and the workplace.

Following the invention of the steam engine, industrialization, and the beginning of the age of computers, we are now in the midst of the fourth industrial revolution with the Internet of Things and Services. The ongoing development of information and communications technology (ICT) has ensured that in large sections of industry, powerful and competitively priced embedded systems, sensors and actuators are now available. Industry 4.0 is the current buzzword used to describe developments towards a production environment that consists of intelligent and autonomous objects temporarily and deliberately networking with one another to carry out tasks. Cyber-physical systems (CPS) and cyber-physical production systems (CPPS) are additional features spoken about in this context [11][15]. CPSs are systems that link

Fig. 11.1 The work of the future – between people, technology and business (Fraunhofer IAO/Shutterstock)

the real and virtual worlds within an Internet of things, data, and services. A broad field of applications is beginning to emerge in the areas of automation, production and robotics, for example, as well as in healthcare and energy distribution [1][9][5].

Successful development and integration of digital technologies within processes in the industrial application sectors is key for Germany's competitiveness [7]. This entails identifying successful responses to new challenges: how can we harness the opportunities for industry, public administration and society afforded by digitization, and how can we overcome the challenges together? How do we want to live, learn, and work in a digital world? How can the possibilities provided by new technologies be reconciled with the demands of demographic change and work-life balance? How can the competitiveness of companies and at the same time the quality of work be positively influenced and continue to increase?

11.2 Future Work Frame – Developing the framework for sustainable work design

The increasing use of digital technologies in production and related fields of work is giving rise to new forms of socio-technical working systems, leading to massive changes in the organization and structure of work. Volatile markets as control mechanisms of our work are leading to an increasing degree of flexibility in terms of time and space. The mobility requirements on employees grow ever greater; new forms of employment are increasingly making inroads alongside regular working relationships. Training in digitization, IT, and intelligent technical systems is ever more becoming the "entry ticket" an employee requires for numerous roles [3][15]. In future, human-technology interaction, workplace flexibility, and the necessary competency and training requirements will need to be actively taken into account when designing work.

11.2.1 Human-technology interaction

The increasing autonomy and intelligence of technological systems is changing the requirements for human-technology interaction. Today, there is broad agreement that the full potential of Industry 4.0 can only be harnessed via people and technology working together in partnership. Human-technology interfaces will in future be of prime importance. They need to facilitate close cooperation between people and technology so that the strengths of technology – such as repeatability, precision and endurance – and unique human abilities such as creativity and flexibility complement one another optimally.

For autonomous and self-learning/self-optimizing systems in particular, there are currently few solid findings regarding how human-technology interaction can be designed to both achieve technological and economic goals while also creating people-centered working conditions that encourage contentment and personal growth [1].

11.2.2 Flexibility, blurred boundaries, and work-life balance

In spite of all the changes brought about by digitization and continuing automation, future work systems in the office and factory will still nevertheless remain socio-technical systems. In this flexible and connected environment, staff members will take on different roles. Their cognitive abilities will enable them to close the sensory gaps of technology, grasping complex situations quickly and comprehensively. As decision-makers, they will resolve the conflicts between networked objects and use digital tools to intervene in time-critical processes. As actors in their own right, employees will complete irregular and highly complex tasks. In their role as innovators and process optimizers, they will continue to be actively involved in the further development of industrial value creation in the future [15]. Equipped with mobile devices, staff will thus work on different tasks with a high degree of flexibility in terms of time, space, and content matter. The scope of future jobs will shift between executive and supervisory or management activities. Mobile forms and content of work will supplement the existing structure of work. In the manufacturing sector, too, the degree of flexibility will increase such that even in production, flexible working locations and flextime will in future become relevant topics.

Alongside massive increases in market-side requirements for flexible staffing, from the employee side new demands for flexible working have also grown in recent years. Here the requirement for temporary and self-determined changes to working hours because of concerns for a healthy work-life balance, for empowerment, and the trend towards self-management are especially at the fore[13][4][12].

11.2.3 Competency development and qualification

For digitization of production and production-related fields to be sustainable, specialist technical competencies in mechanics, electrical engineering, microtechnology, IT, and their combinations are already required today. Further, a deeper understanding of the physical and digital processes and how to synchronize them in near-real time is also needed [16]. Competencies for cross-discipline and cross-process communication, cooperation and organization are becoming indispensable for work in interdisciplinary teams and networks.

As production systems are enhanced with Industry 4.0 techniques, constant interaction with continuous innovation and change is becoming the norm due to the requirements of a flexible production process and of continuous technological and lasting technological-organizational adjustments. In order to successfully master these changes, there is a need for widespread competency development to provide qualification for Industry 4.0. This means unlocking and developing the ability of each and every staff member to develop their own know-how and knowledge [6][14].

Competency – the combination of knowledge and know-how – is best developed by staff learning to carry out tasks that they had not mastered before during the specific operational working process [8]. Accordingly, qualification for Industry 4.0 should be at the heart of "on-the-job" competency development/learning oriented around the work process. Here, work tasks would be supplemented by learning tasks, for example. Learning would be supported via advice and coaching concepts and, supplemented by new digital tools, may be either self-managed or take place in groups.

11.3 Future Work Trends – Work design in Industry 4.0

The enhanced features of CPS provide the opportunity to redesign the industrial work of the future. When it comes to workplace design, the key elements which build upon one other on the user side are networking, context sensitivity, assistance, and intuitivity. Intelligent workplaces and work systems will be enhanced by digitized work organization. In what follows, we will go into more detail in the workplace level only.

11.3.1 Connected work systems

The goal of connected workplaces is to connect systems and data from products and processes thoroughly instead of using existing IT silo systems. To do this, workplaces need to be equipped with sensors that capture the accruing data. The transmission,

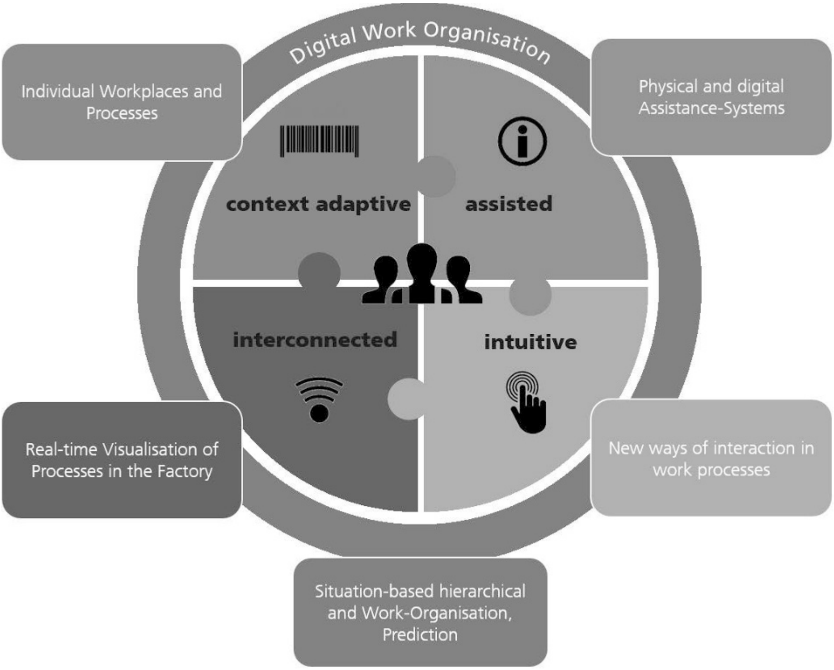


Fig. 11.2 Future Work Trends change work design. (Fraunhofer IAO/Fotolia)

processing and sharing of data between objects in production must also be facilitated. Connected data sharing here should take place both horizontally along the process chain as well as vertically within the company, and bi-directionally – between systems and users, and between users and systems. It is only in this way that the information gained from the data can be made available to systems and staff and utilized by them.

In an Industry 4.0 environment, connected objects and real-time utilization of relevant production parameters find application, for example, when production events are fed back into the production control system (e.g. for incident management). Within the field of connectedness, the design tasks are developing new Production 4.0 concepts (beyond lean) and supporting staff acceptance of work in transparently generating work systems. Furthermore, staff must be qualified for work within highly connected systems and must be able to interpret accruing data in order to optimize their work processes.

11.3.2 Context sensitive work systems

Traditional production structures only take limited account of the diversity of employees and operations, as well as their requirements in the work process. The high number of variants in production makes this necessary however. Context sensitive work systems allow this challenge to be met and form the basis for the vision of batch size 1 production. For the production context to be systematically integrated, the system, or the information to be provided, must be adapted to specific changing environmental conditions. The system continuously monitors the work situation and also knows its user. It identifies the current process sequence by comparing it with (de)centrally stored data and provides the user with data related to the current process step in personalized form.

In an Industry 4.0 environment, the production context can be utilized for workspace personalization, for example by adapting working heights, lighting, and providing operation-specific information to the employee. Defining rules for personal data use and identifying the usable personalization space within which work can be designed to suit the employee or process are current design focus areas.

11.3.3 Assisting work systems

Assisting work systems represent the next stage of development. These systems primarily provide support in mastering the extensive diversity arising from the increasing individualization of products and processes. The following can be identified as requirements for assistance systems [17]:

- Access to the data of intelligent objects, via RFID or other tracking technologies for example, must be ensured.
- The efficient and effortless integration of assistance systems into the working environment and context sensitive information provision support productive work processes.
- The networking of system components to facilitate exchange with centrally held or other decentralized data or to initiate the assistance system's actuator subsystem (e.g. requesting a mobile robot on the basis of work progress).
- Autonomous assistance to facilitate maximum independent decision-making and avoid slowing down the worker in the course of their work.

Today, a broad spectrum of applications for assisted work systems is already conceivable. A distinction needs to be made here between digital assistance systems (e.g. augmented reality/virtual reality) and physical ones (e.g. lightweight robots, exoskeletons). These systems are able to use feedback to adaptively support employees' learning during the process. Planning the distribution of control between the person and the technology here is the key determining factor of their successful introduction.

11.3.4 Intuitive work systems

More and more production tasks are being supported by IT systems or carried out by them entirely. For staff, the result is that these process accompanying systems are becoming increasingly complex to use. Intuitive digital work system design and utilization thus represents an important lever for increasing efficiency. Ergonomic design of the human-technology interaction both physically as well as with respect to information ergonomics is a key requirement here. Control systems for the worker also need to be reduced to process-relevant parameters and override values. In this way, intuitive interaction concepts based on gestures, speech, touch and in future even brain function, combined with mobile devices and wearables, will be able to provide productive support to the work process. For these enablers to be properly utilized, the necessary competencies for managing new tools and interaction concepts need to be identified and formulated – combined with new mobile device use concepts such as "bring your own device" (BYOD).

11.4 Future Work Lab – Experiencing the industrial work of the future

The variety of opportunities, enablers, and new demands for successful work design in industry must be made adaptable in a practical environment and tangible for all stakeholders. In the Future Work Lab, a new innovation laboratory for work, people, and technology is being developed in Stuttgart, dedicated to these issues. The Future Work Lab functions as an interactive shop window and ideas center for sustainable and people-centered work design in production and related fields. In the Future Work Lab, the design of future industrial work in Germany is being discussed, participatively advanced, and made tangible in close coordination with the relevant parties. The foundation for this is already being laid via pilot and soon-to-be-implemented digitization and automation solutions.

In order to achieve this, the Future Work Lab is divided into three areas [10]:

- A central demo world with *Workplace 4.0 Tracks* that introduce visitors to the workplace of the future via hands-on exhibits of different forms of digitization and automation,

Fig. 11.3 Structure of the Future Work Lab (Fraunhofer IAO/Ludmilla Parsyak)

- The *Fit for Future Work* learning world providing information, qualification and discussion opportunities on the developmental trends of future workplaces
- The *Work in Progress* world of ideas, a think tank for research into work and a safe space for the design and development of new and unforeseen or hard-to-foresee solutions

The close relationship between these three areas ensures that the latest solutions for the work of the future can constantly be discussed, developed and demonstrated in an open and adaptable environment on the basis of current solutions in the context of technological transfer.

11.4.1 Experience Future Work demo world

The *Workplace 4.0 Tracks* – demonstrators with content related to one another in the demo world – are designed to tangibly show short- and medium-term changes in the workplace as well as long-term developmental trends. For this, 50 different demonstrators are being put together in the Future Work Lab. The layout is oriented around today's typical work profiles across the operational value chain.

The *Today+* track shows operational use cases demonstrating industrial work over the period from 2016 to 2018. The Future Work Lab thus illustrates developments within the industrialized and modern medium-sized enterprise – lean production, lean systems, and integrated production systems.

The long-term developmental trends demonstrate operational use cases for the digitization and intelligent automation of industrial work of the time horizon up to 2025. They illustrate diverse demonstrators at the crossroads between technology-centered automation and human-centered specialization that may be standard in the manufacturing industry in 2025. In this way, both developmental scenarios of the Industry 4.0 currently discussed are referred to: the automation and specialization scenarios [16]. In order to illustrate these, potential Industry 4.0 operational use cases are combined: it is possible to demonstrate and experience how work in the future can on the one hand be designed to be more technology-oriented and on the other hand more strongly people-focused. In the Future Work Lab, different developmental trends and their consequences are highlighted in this way as they relate to work design, technology integration, competency or qualification requirements, and so on.

The available demonstrators serve as the basis for the Future Work Lab. They serve the other elements of the laboratory – the learning world and world of ideas – as an interactive work, learning and research environment.

11.4.2 Fit for the Work of the Future learning world

Industry 4.0 also entails working on (semi-)automated equipment and in virtual environments. At present, for medium-size enterprises in particular, these kinds of solutions are generally neither available physically nor as virtual simulations for process-oriented competency development, at least in the state of the development envisaged. In the learning world, the Future Work Lab demonstrators are utilized for competency development as well as for the shared discussion, development, and testing of suitable qualification concepts. In addition, the implications for competency development within the *Work in Progress* world of ideas for research into work are identified and addressed in good time.

The competency development and advice center concept extends beyond the notion of a learning factory. Alongside the opportunity to make technological applications tangible and learnable via the demonstrators, the following are also on offer:

- Support in recognizing the learnability and work quality of Industry 4.0 applications.
- Conducting "future workshops" together with medium-size companies in order to illustrate future development scenarios of Industry 4.0 applications via specific company examples, and in order to analyze changing tasks, requirements, competencies and potential competency development paths.
- Modelling business and work processes that arise from Industry 4.0 applications and demonstrating them using a virtual reality platform.
- Training and learning modules for modern methods of participative design of interactive, adaptable production systems.

The formats developed in this part of the laboratory enable different target groups (e.g. management, planners, team leaders, works committee members, and employees) to be informed and advised regarding design options and to develop them participatively.

11.4.3 Work in Progress world of ideas

The *Work in Progress* world of ideas reinforces the academic character of the Future Work Lab, establishing a think tank as a safe space for research, innovation and dialog around the future of work and the human-technology interaction it entails. The world of ideas thus stands for sharing about and developing new solutions for the workplace of the future based on cyber-physical systems.

A central theme of the center's academic orientation is developing a descriptive model for industrial work in the course of the digital transformation. The model is being worked out in combination with the available human-technology design options identified during demonstrator development, and validated by means of user testing with the demonstrators.

In the *Monitoring and Benchmarking section*, academic exchange in the field of work research is facilitated. A world map of work research provides an overview of national and international research findings.

11.5 Future Work Cases – Design examples for the industrial work of the future

When complete, the Future Work Lab will offer more than 50 different demonstrators: they will show what the changes in work prompted by Industry 4.0 might look like across the value chain. The demonstrators are being produced for operational fields of work such as machine operation, assembly, factory logistics with receiving and shipping departments, quality assurance, scheduling, maintenance, and industrial engineering. The first demonstrators implemented can be seen in the illustrations. In what follows, we will take a closer look at two Future Work Cases by way of example.

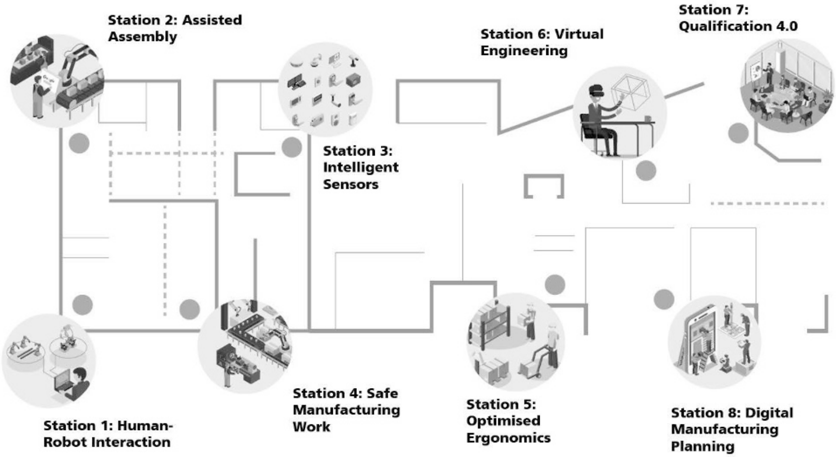


Fig. 11.4 Demonstrators in the Future Work Lab (excerpt) (Fraunhofer IAO/Fotolia)

11.5.1 Future Work Case: assisted assembly

Customer-driven markets require companies to provide multi-variant product portfolios as well as structures equipped for batch size 1 production. For staff in assembly, this poses significant demands in terms of constantly changing and complex tasks. In the context of digital assistance systems, different technologies may be combined in order to train staff quickly and intuitively or guide staff directly in the assembly process and capture information on the production process.

The Future Work Lab on the one hand features demonstrators that support new staff in learning complex work processes via training videos. The videos are used both in small animated segments as repeatable qualification units (“knowledge

Fig. 11.5 Assisted assembly work in the Future Work Lab (Fraunhofer IAO/Ludmilla Parsyak)

nuggets") as well as on demand, and touching different process-relevant topics. Staff can thus expand their knowledge independently and in a decentralized manner as the need arises or during idle periods. Systematic progress through the learning process here is governed by the digital assistant and staff motivation is simultaneously increased.

Also on display in the Future Work Lab are assembly stations that use digital assistants to guide workers through the assembly process. Concepts in use here include the beam projection of process-specific information, pick-to-light systems for material and tool picking, or process videos that show how specific operations should be carried out correctly. These assistants are enhanced via range or Kinect cameras that record workers' movements and make the digital assistance system accessible. Localization technologies such as ultrasound additionally enable the positions of workers or tools to be identified.

Digital assembly assistance tools actively reduce information complexity for workers by displaying context-sensitive, situation-based and personalized information. In this way, staff receive feedback on their assembly work in real time. Quality assurance thus takes place directly within the process. This increases the efficiency of operation as well as the quality of output and staffing flexibility.

11.5.2 Future Work Case: human-robot cooperation with the heavy-duty robot

Robotic systems have been taking on a central role in industrial work for several decades now – a role that will further expand significantly in future. New robotic systems – in particular lightweight robots as well as improved safety technologies and correspondingly adapted operational processes – ensure that robots are more and more becoming the physical assistants of humans in a range of different processes.

One of the demonstrators in the Future Work Lab shows what this kind of workplace for human-robot collaboration (HRC) might look like. The unique feature here is that it shows that HRC is not only achievable with lightweight robots. While these small, compact systems with partially integrated force/torque sensors are inherently safer than heavy-duty robots, their load-bearing capacities are equally limited by virtue of their construction.

In the aforementioned workplace, the workspace is open, and the worker can directly monitor and coordinate the progress of the work. In this way, the specialist knowledge and dexterity of the individual can be combined with the strength and endurance of the robot. The results are workplaces with improved ergonomics and

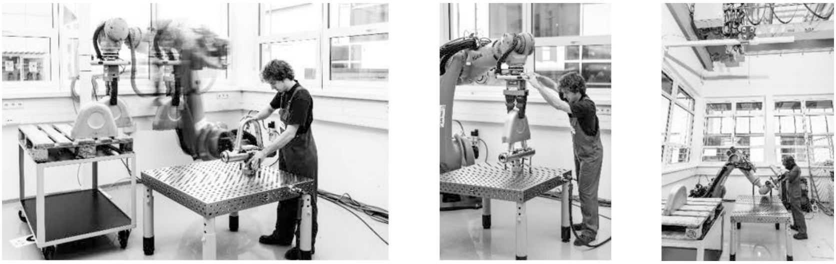


Fig. 11.6 Human-heavy-duty robot collaboration in the Future Work Lab (Fraunhofer IPA/Rainer Bez)

higher productivity and quality. This is made possible by the workplace safety-certified SafetyEye camera system that watches over the robot's working space from above. The system recognizes when people are approaching the robot's working space. The robot then either reduces its speed or stops altogether in order to guarantee the individual's safety. The robot can also be switched to a manual operation mode.

Human-robot collaboration thus not only offers benefits by virtue of improved ergonomics due to the robot taking over physically demanding tasks. Specific quality-critical processes can also be safely carried out by robots. In addition, the scalability and personalization of production also become increasingly important in the Industry 4.0 context. Since industrial robots can be employed universally, they generally provide good conditions for implementing versatile production.

This is all the more the case if the "rigid monuments" of worker safety railings that have thus far been so common in factories are in future done away with.

11.6 Outlook

Following the construction of the Future Work Lab, the innovation laboratory will enter its launch and operational phase, to be designed in cooperation with its users, for example its social partners. In addition, the more than 50 demonstrators will continue to be actively developed and incorporated into new research assignments. The training and advice formats will receive new impetus based on constant use, and research on the demo world will permit new scientific findings, in the context of user experience for example.

Collaboration with companies, associations, labor unions, and employees will ensure developments here are practical and will facilitate the access to and transfer of research outcomes for all the stakeholders involved.

Internationalization will be key to the Future Work Lab's ongoing establishment as a lighthouse for application-relevant work research. Networking and exchange with other innovation laboratories, researchers, start-ups and actors globally will enable the early incorporation of trends into the Future Work Lab as well as international positioning for the ideas developed.

By designing sustainable working environments participatively, the Future Work Lab is making a significant contribution to increasing companies' competitiveness long term.

Sources and literature

[1] acatech; Forschungsunion (Hrsg.) (2013): Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0. Abschlussbericht des Arbeitskreises Industrie 4.0. https://www.bmbf.de/files/Umsetzungsempfehlungen_Industrie4_0.pdf. Zugegriffen: 27.02.2017.

[2] acatech (Hrsg.) (2015): Innovationspotenziale der Mensch-Technik-Interaktion. Dossier für den 3. Innovationsdialog in der 18. Legislaturperiode. Berlin. http://innovationsdialog.acatech.de/themen/ innovationspotenziale-der-mensch-maschine-interaktion.html. Zugegriffen: 19.05.2017.

[3] Bauer, W. et. al. (2014): Industrie 4.0 – Volkswirtschaftliches Potenzial für Deutschland, https://www.bitkom.org/noindex/Publikationen/2014/Studien/Studie-Industrie-4-0-Volkswirtschaftliches-Potenzial-fuer-Deutschland/Studie-Industrie-40.pdf. Zugegriffen: 15.05.2017.

[4] Bauer, W.; Gerlach, S. (Hrsg.) (2015): Selbstorganisierte Kapazitätsflexibilität in Cyber-Physical-Systems. Stuttgart: Fraunhofer Verlag.

[5] Bauernhansl, T., Hompel, M. ten u. Vogel-Heuser, B. (Hrsg.) (2014): Industrie 4.0 in Produktion, Automatisierung und Logistik. Anwendung, Technologien, Migration. Wiesbaden: Springer Vieweg.

[6] Bergmann, B.; Fritsch, A.; Göpfert, P.; Richter, F.; Wardanjan, B/Wilczek, S. (2000): Kompetenzentwicklung und Berufsarbeit. Waxmann, Münster.

[7] BMBF Bundesministerium für Bildung und Forschung (2014): Die neue Hightech- Strategie. Berlin, https://www.bmbf.de/pub_hts/HTS_Broschure_Web.pdf. Zugegrif- fen: 27.03.2016.

[8] Bremer, R. (2005): Lernen in Arbeitsprozessen – Kompetenzentwicklung. In: Rauner, F. (Hrsg.): Handbuch Berufsbildungsforschung. wbv, Bielefeld.

[9] Broy, M. (Hrsg.) (2010): Cyber-Physical Systems. Innovation durch Software-intensive eingebettete Systeme. Heidelberg: Springer.

[10] FutureWorkLab (2017): Fraunhofer IAO, www.futureworklab.de. Zugegriffen: 17.05.2017.

[11] Ingenics und Fraunhofer IAO (2014): Industrie 4.0 – Eine Revolution in der Arbeitsgestaltung https://www.ingenics.de/assets/downloads/de/Industrie40_Studie_Ingenics_IAO_VM.pdf, Stuttgart. Zugegriffen: 17.05.2017.

[12] Hämmerle, M. (2015): Methode zur strategischen Dimensionierung der Personalflexibilität in der Produktion. Wirkungsbewertung von Instrumenten zur Flexibilisierung der Personalkapazität im volatilen Marktumfeld. Fraunhofer Verlag Stuttgart 2015.

[13] IG Metall (2013): Arbeit: sicher und fair! Die Befragung. Ergebnisse, Zahlen, Fakten. https://www.igmetall.de/docs_13_6_18_Ergebnis_Befragung_final_51c49e134f92b4922b442d7ee4a00465d8c15626.pdf. Zugegriffen: 19.05.2017.

[14] Röben, P. (2005): Kompetenz- und Expertiseforschung. In: Rauner, F. (Hrsg.): Handbuch Berufsbildungsforschung. wbv, Bielefeld.

[15] Spath, D. et. al. (2014): Produktionsarbeit der Zukunft – Industrie 4.0. FhG IAO, http://www.produktionsarbeit.de/content/dam/produktionsarbeit/de/documents/Fraunhofer-IAO-Studie_Produktionsarbeit_der_Zukunft-Industrie_4_0.pdf. Zugegriffen: 17.05.2017.

[16] Spath, D.; Dworschak, B.; Zaiser, H.; Kremer, D. (2015): Kompetenzentwicklung in der Industrie 4.0. In: Meier, H. (Hrsg.): Lehren und Lernen für die moderne Arbeitswelt. GITO Verlag, Berlin.

[17] Wölfle, M. (2014): Kontextsensitive Arbeitsassistenzsysteme zur Informationsbereitstellung in der Intralogistik. München TUM, Zugleich Dissertation, München, Technische Universität München, 2014.

Cyber-Physical Systems

12

Research for the digital factory

Prof. D.Eng. Welf-Guntram Drossel ·
Prof. D.Eng. Steffen Ihlenfeldt · D.Eng. Tino Langer
Fraunhofer Institute for Machine Tools and Forming Technology IWU
Prof. D.Eng. Roman Dumitrescu
Fraunhofer Institute for Mechatronic Systems Design IEM

Summary

Digitization is the defining innovation driver for value creation in the modern global industrial society. At the forefront stand the increase in efficiency for flexibilization and improved resource utilization provided by the self-optimizing automation of processes. Digital technologies must become inherent components of the production system.

A cyber-physical system represents the sought-after unity of reality and its digital reproduction, and is the next stage in development of mechatronics into a symbiotic systems approach based on the IT networking of all components. IT together with non-technical disciplines have produced a range of methods, techniques and processes by which sensors, actuators, and cognition can be integrated into technical systems so they demonstrate functionalities that have only been fulfilled by biological systems until now. In this way, the evolution prompted by Industry 4.0 technologies leads to a genuinely disruptive paradigm change. Production, suppliers, and product developers enter a new quality of innovative cooperation.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_12

12.1 Introduction

Digitization is the defining driver of innovation for value creation in the modern global industrial society. This is the reason for a range of activities that are often grouped together today under the heading of *Industry 4.0* or *IoT – Internet of Things*. All of these approaches have in common that they prioritize efficiency gains required for enhancing flexibility and improving utilization of resources in production, via the self-optimizing automation of processes.

Nevertheless, connecting production with the latest IT and communications technology via Internet technologies – often described as the *fourth industrial revolution* –poses enormous challenges for numerous companies. In addition to the technical controllability of much more flexible production and supplier networks, it also entails a far-reaching economic revolution. The enhanced technological possibilities also come with a change in traditional customer-supplier relationships as well as global market access. The global flexibilization within supplier networks means a disruption of the usual sharing of risk in traditional supplier chains. For the supplier industry in particular, which is characterized by small and mid-size enterprises, this in turn holds huge economic risks.

New opportunities, and in particular new potentials for employment, will only arise as results if digitization becomes a business model not only for developers and suppliers of software. The traditional sectors of German industry, such as machinery

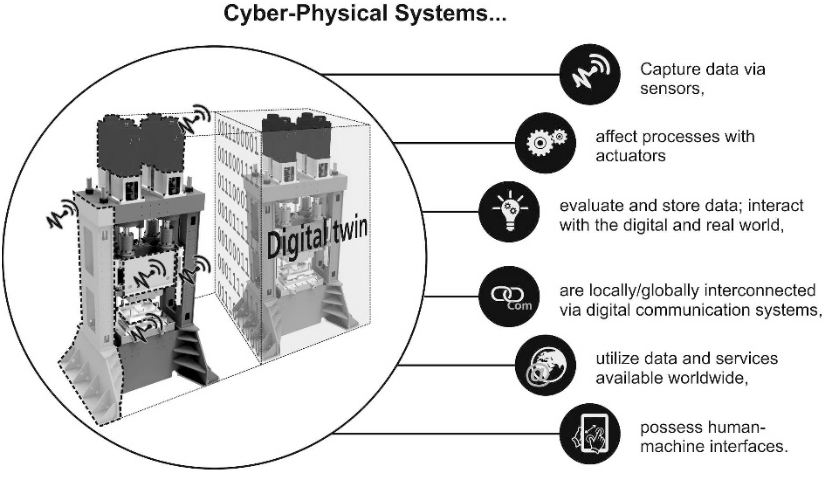


Fig. 12.1 Definition of cyber-physical systems (Fraunhofer IWU)

and plant engineering, and automobile manufacturing, have to obtain the capability of using digital technologies to produce new products and services, which also implies new business models. Digital technologies have to become an intrinsic component of the production system and production facility.

A cyber-physical system (CPS, see Fig. 12.1) represents the sought-after unity of reality and its digital reproduction – a further stage in the development of synergetic approaches to mechatronics (combining the best of all disciplines) towards a symbiotic systems approach based on the IT networking of all components.

By now CPSs have become a key trend in product development. Initial applications include, for example, intelligent electricity meters in smart homes or self-orienting logistics systems in smart factories. Autonomous vehicles are another example of a future system that will be based on these same principles.

The information technology and non-technical disciplines such as cognitive science or neurobiology have given rise to a range of methods, techniques, and processes by which sensors, actuators, and cognition can be integrated into technical components. These components then demonstrate functionalities that have only been fulfilled by biological systems thus far. As a result, CPSs are significantly more than connected mechatronic structures. They provide the basis for fascinating perspectives on technological systems [1]:

1. Autonomous systems: they solve complex tasks independently within a specific application domain. They must be in the position to act productively without remote control or other human assistance. For example, the basis of actuator control may be based on an environmental model within the system, enabling it to learn new events and new actions during operation. This requires a number of technological building blocks such as sensor fusion, semantic explanatory models, or planning processes, for example [2].
2. Dynamically connected systems: the degree of system networking will increase. This will lead to new and increasingly complex systems whose functionality and capacity exceed the sum of the individual parts. The system boundaries, interfaces, and roles of the individual systems vary depending on the goal of the system as a whole. The networked system, which increasingly functions as a unified whole, will no longer be exclusively controllable by means of global control, rather the desired global behavior will have to be achieved via local strategies. One example of this is the light-based navigation of driverless transport systems that only becomes possible via the interaction of numerous individual systems. However, they can function independently of one another and are developed either independently or by a number of different suppliers [3]. For this reason the term *system of systems* (SoS) is used [4].

3. Interactive sociotechnical systems: the outlined path of technological development also opens up new perspectives on the interaction between humans and machines. The systems will adapt flexibly to user needs, providing context sensitive support. In addition, they will also be capable of explaining themselves and providing the user with possible actions. Interaction will increasingly become multimodal (e.g. speech- or gesture-based), and take place via a diverse range of technologies (e.g. augmented reality or holograms). The result is a complete sociotechnical system [5]. Against this backdrop, the question will be less which tasks people will be replaced for, but which new or existing tasks can be solved in a new way by using augmentation.
4. Product/service systems: the continuing technological development of systems will not only change engineering but also the entire market offering. Product/service systems will be developed based on the close interlinking of physical and service offerings and they will provide customized solutions to problems. Data-based services that incorporate the collection, processing, and analysis of data are the main source of the benefit from these kinds of new solutions. Data analysis (e.g. a prognosis of likely imminent machine failure and preventative maintenance) can be used for offering tailored services (e.g. automatic ordering of replacement parts) [6]. Smart combinations of innovative services and intelligent systems form the basis for innovative business models [7].

12.2 CPSs in production

Making these kinds of scenario projections for manufacturing is unique since manufacturing is not a technically homogeneous system. Alongside an enormous technological diversity, there are also large variations in how technically advanced production facilities are, together with a great range of organizational forms. One key reason for this, alongside economic boundary conditions such as business size or position in the value chain, lies in the extreme variation of innovation and investment cycles. The lifecycle of production equipment, for example, may range from several years to several decades, while the innovation cycle in the software industry often equates to just a few weeks.

For this reason, it is not only the design of cyber-physical systems that is of paramount importance, but also the development of methods for transforming the structure of production systems into the CPS system architecture. The migration of existing production plants is only possible through the implementation of intelligent, connected subsystems. Their collaboration can only be guaranteed, if communication between all of the subsystems can be ensured. However, the basic

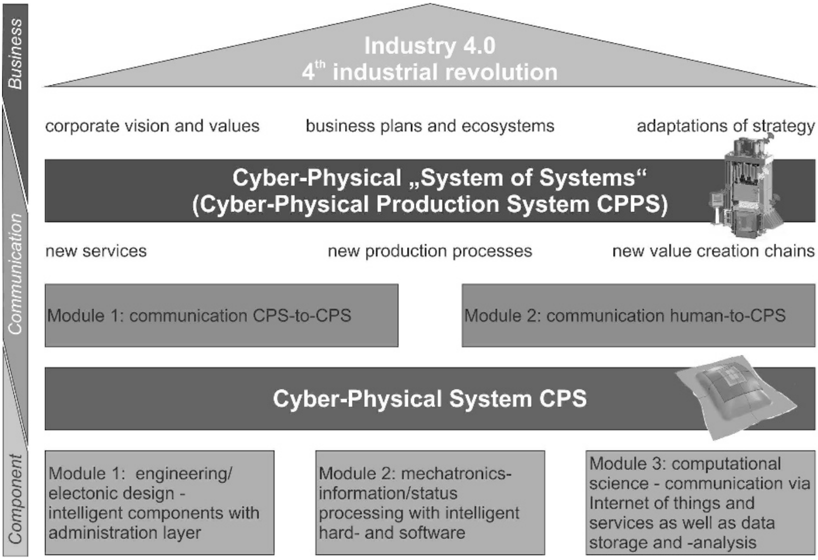


Fig. 12.2 Schematic classification of CPSs and CPPSs (according to [9])

structure of production systems will change very little in the near future. Components such as sensors, drive units, or frames will largely retain their respective core functionality, but will require software components with functional and structural models as well as communications hardware and software in order to upgrade them into CPSs. As shown in Fig. 12.2, three domains can be distinguished at the basic level:

- CPSs for mechanical engineering and electrical engineering include, for example, intelligent machine frames with integrated sensors (e.g. to measure force) in order to send measured data to a superior monitoring level. An additional example could be integrated sensor nodes that measure temperatures and accelerations, combine them on single-board computers, and pre-process them.
- CPSs for mechatronics include, for example, integrated active attenuators with built-in sensors, actuators, computing unit and the ability to communicate.
- The third aspect at the component level is CPSs for IT/data, with the Internet of Things (IoT) being the main enabler for the remaining domains. These store recorded data and calculate simulation models and digital reproductions of systems and system components, either using master computers near the machines or applying cloud systems.

In keeping with the CPS definition [8], the components of a system are connected via the hierarchy levels (e.g. the automation pyramid). Thus, the communications level in Fig. 12.2 gains a particular significance. Once again, two classes of communications are distinguished here: communications between CPSs (according to specific functionality) and between CPSs and people, e.g. via control panels, smart glasses, cellphones, or tablets.

Further, these production systems also permit enhanced value creation since it is no longer merely the manufactured products that represent value for the customer, but increasingly also the data recorded. This is a key benefit of intelligent connected systems, for the automotive and aerospace industries in particular, due to their greater requirements for documentation.

Cyber-physical systems from various domains can be combined into an overarching system ("system of systems"). For production systems use cases, they become cyber-physical production systems (CPPS). These offer potential for

- New business models and ecosystems such as leasing models and tailored provision of certain functions for processing units,
- A changing orientation for companies towards information- and data-driven service products, e.g. by integrating service planning and predictive maintenance for their own products as well as
- Customized production and process control based on additional information gained from the process, and the knowledge thus generated.

For these reasons, manufacturers and users of cyber-physical production systems do not only have to adapt to new machine generations with an increased range of functions, but also to new perspectives and opportunities for the orientation of companies and value-added networks.

12.3 Transforming production systems into cyber-physical systems

12.3.1 Evolution in the production process

As part of the Deutsche Forschungsgemeinschaft's (DFG) Collaborative Research Center 639, the TU Dresden Institute of Machine Tools and Control Engineering carried out the evolution from a conventional production system into a cyber-physical production system, in particular in terms of a data-driven process map [10][11], and tested it on an example process of "spring dome manufacturing".

This example process comprises the following steps:

- Inserting the preform into a variotherm tool,
- Closing the forming machine and maintaining the pressing force while heating (integrated temperature measurement),
- Opening the tool after a defined dwell time, removing the component, and
- Inspecting the component and assessing its quality characteristics.

In this case it is essential that the melting temperature of the thermoplastic material is achieved and the consolidation process started while the tool is closed. The entire process is characterized by a strong interaction between the properties of the initial materials and the combination of the process parameters used for consolidation. Inferring corrective actions for achieving good parts under changing boundary conditions is thus not trivial, and the process execution is comparatively complex.

An overarching goal when expanding a production system into a CPPS is being able to monitor manufacturing processes and, in the case of parameter variance, being able to influence them in such a way that good parts are still produced. The essential elements here are data acquisition, modeling, and feedback to the machine control (cf. Fig. 12.3).

The following steps are taken in order to acquire the necessary process information:

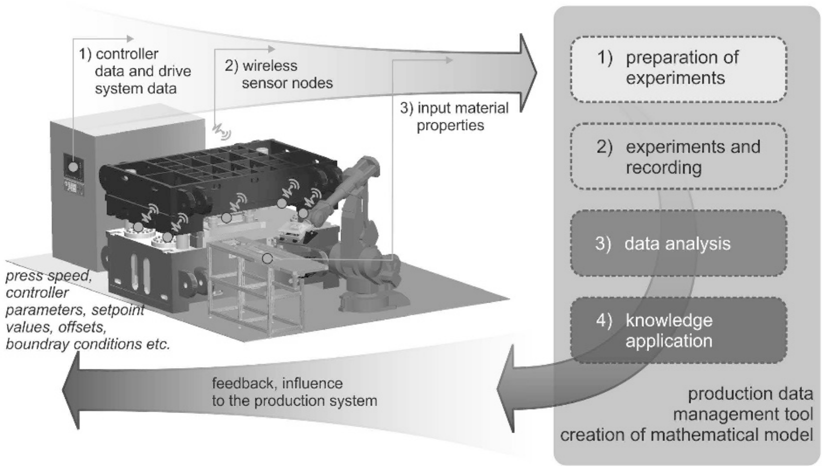


Fig. 12.3 Production system with extensions necessary for being turned into a CPPS (with elements from IWM, TU Dresden)

1. Data is acquired from the machine control (SPS, CNC, motion control) and from the drive systems of all of the actuators,
2. Additional sensors are attached to the production system and to the utilized tools; their signals are also recorded,
3. The properties of the semi-finished product are identified – here, application-specific sensor units often have to be developed and installed.

Information from the drive systems can be acquired via software extensions of the machine control system (in this specific case, a CNC) and via the fieldbus [12]. The data transmission from the control to a management system may take place by means of the OPC unified architecture (OPC UA) interface, for example, something which current observations suggest is crystallizing into an unofficial standard for Industry 4.0 [13].

Once the data is available, it can be processed and analyzed using a process data management tool. The goal is the representation of the process in the form of a model (mathematical model, black box model, or system simulation) which can be used for carrying out extensive analyses. This requires a decentralized knowledge base for material and component properties together with the interactions in question, available in cyberspace. It is then used to identify the optimum process parameters that ensure a stable and reproducible process according to current input criteria and boundary conditions. An important aspect in developing the knowledge base is a procedure for process description and experimental design that is as efficient as possible. Fig. 12.4 illustrates the individual steps from process description through to its reconfiguration.

For manufacturing the spring dome, a graphical process description (input, processing, output, cf. Fig. 12.5) was found to be highly suitable. This encompasses

The steps of pre-shaping, sensor placement, consolidation, and component quality assessment. The interrelationships and interactions between the following aspects are registered and mathematically modeled:

- The properties of the initial materials,

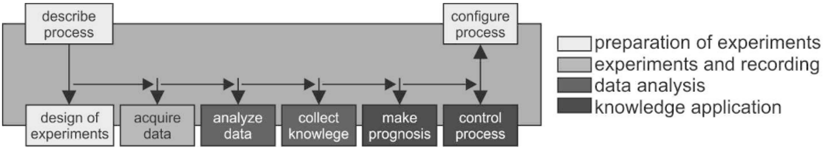


Fig. 12.4 Modelling procedure (IWM TU Dresden)

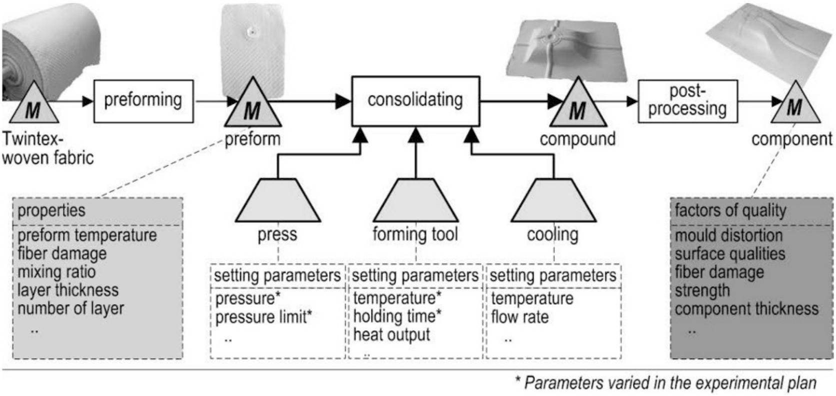
- The set of parameters for the production system,
- The environmental impact parameters, and
- The desired quality of the parts manufactured [10][11].

Based on the graphical process description and the registered interrelationships, the mathematical model is able to identify situational optimal process parameters for the production system. Thus it ensures that production systems have the desired greater adaptability. This implies that it represents a core functionality of cyber-physical production systems.

Based on the registered graphical process model, the necessary experiments for identifying the required characteristic values are calculated via statistical design of experiments.

In this specific case, a D-optimal design was selected since this permits continually changing variables (e.g. temperature) and discrete states (true/false) to be equally incorporated into the planning process. The main influencing parameters from Fig. 12.5 were systematically varied and corresponding experiments carried out. The respective results were logged in terms of component quality and recorded in the mathematical model.

Based on the model, it was possible to test various process settings and to assess their effects on component quality. In doing so it is possible to map destabilizing settings of the process. Beginning with a desired level of quality for the parts produced, it is also possible to deduce the required parameterization of the process/machine. This represents a basis for auto-tracking of the control parameterization. A cloud solution is best for calculating in a parallelized manner variant analyses and



* Parameters varied in the experimental plan

Fig. 12.5 Graphical process record (IWM TU Dresden

Table 12.1 Influencing parameters of the manufacturing process

Properties of the preform	Process parameters	Quality characteristics of the components
Temperature	Temperature	Component thickness
Proportions of the individual components	Pressure during consolidation	Stability of the parts
Number of layers	Dwell time	Surface quality
Layer thickness	Cycle time	
Pattern of the glass fiber orientation	Maximum energy intake	

calculations of optimization for the interaction between control parameterization and the process model. Users are provided with a graphical interface that can be used to test and assess various parameter combinations using sliders. The mathematical model produced meets the temporal requirements for serving as the source of reference values for feedback to the machine control.

The goal of the data analysis is to identify measures for improving the production process. This results in changes to the production system itself or to the machine control. This may be achieved by means of additional actuator systems, or it may be restricted to adjusted parameters of the base system. In the example chosen, the feedback comprises process-relevant parameters such as the stroke rate of a press, dwell times, and press forces. The feedback also requires communication between the control level, which features the integrated process model, and the machine control (cf. Fig. 12.3).

Current controls allow a range of their parameters to be adapted in the control cycle. As with data acquisition, this requires the realization of a communications link with the modeling level. This may be achieved with varying degrees of complexity and integration. Simple solutions are limited to the exchange of current process parameter vectors via text files or specific shared memory areas. It is also possible to envisage process models and machine control being linked via fieldbuses. Here, the preferred solution in each case is guided by the control solutions utilized.

An additional aspect is the frequency of parameter adaptation in the control, which implies the cycle of the (quality) control loop developed. Semi-continuous regulation of the process is necessary in the case of large variations in component quality and major dependencies of entry criteria and boundary conditions.. Its calculation cycle is guided by the control cycle times, which typically range from 0.5 to 3 ms [12]. Nevertheless, this entails very high requirements for the process model, which must supply a new set of parameters to the control after each cycle, and must be calculable within the control cycle. Furthermore, it limits the representable model complexity and requires fast communication, e.g. via a fieldbus. More limited variations in component quality and higher time constants in the process can be managed via "part-to-part regulation". This does not require real-time model calculation and communications, but nevertheless poses challenges e.g. for the measurement of initial material parameters. The least demanding regulation strategy allows for a parameter adaptation only after changing the semi-finished product ("batch-based regulation"). Although there is a range of processes with stable behavior throughout a batch of the initial material, they still require re-parameterization after changing the semi-finished product. Due to longer calculation periods possible even very complex process models can be calculated in this case, and alternative considerations carried out in order to achieve suitable parameterization.

Enhanced data acquisition, networking with data transport, and intelligent (sub-) components are all core functions of cyber-physical production systems. Developments in cyber-physical systems can be observed in the various domains. The example selected demonstrated that CPPS can be equipped to react autonomously to changes in the process and ensure production of OK parts by using data acquisition, modeling, and feedback to the machine control.

12.3.2 LinkedFactory – data as a raw material of the future

Cyber-physical production systems will only reach their full potential when data and information are used beyond the flexibilization, control, and quality assurance of the internal technological process. The increasing networking of machines and logistics systems is driving the volume of "data as raw materials". The availability of ever greater data volumes offers huge potential for the targeted analysis of the information contained.

Many companies strive for constant availability of all relevant data and information on key processes and procedures. The goal is to provide details on the current state of production quickly and easily, with a reliable outlook on the near future by means of suitable forecasting approaches where necessary. In order to meet the existing requirements, suitable systems for information and communications are required for recording and providing the relevant data, and in particular the information inferred from it. The heterogeneity of the data sources here requires an appropriate, flexible IT infrastructure in order to later generate knowledge and to

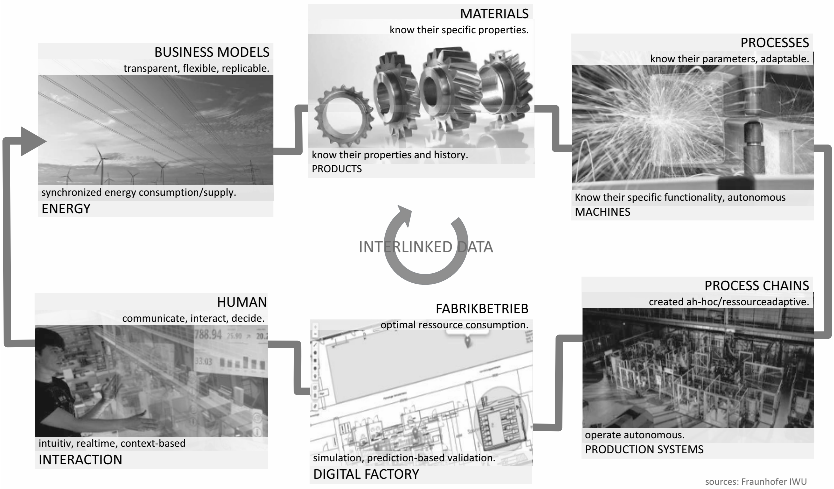


Fig. 12.6 Networked data in the production environment (Fraunhofer IWU)

support decision-making using data and information, using semantics and cognitive algorithms.

The overall system must integrate both hardware and software solutions into production. These solutions have to be flexibly established and decentralized. With each new system, the complexity increases, as does the risk of diminishing technical and indeed organizational availability. For this reason, the focus lies on ensuring a high level of resilience at all levels, from the sensors through communications and IT infrastructure to the selected algorithms in order to achieve the synergistic protection of the overall production system. .

At present, the data captured is often only analyzed and processed in line with its original reason for capture. By correlating data that has previously been managed in individual systems, the use of suitable methods of analysis facilitates the inference of new information [14] (cf. Fig. 12.6). Here, the mass of data should be linked and consolidated in such a way that humans can make the correct decisions in production. This in turn is a precondition for agility and productivity [15].

Fully establishing this kind of approach requires new, modular component solutions for heterogeneous production systems to be made available. In order to achieve this, the following research questions need to be answered, among others:

- How can the heterogeneity of production and information technology be managed and minimized economically?

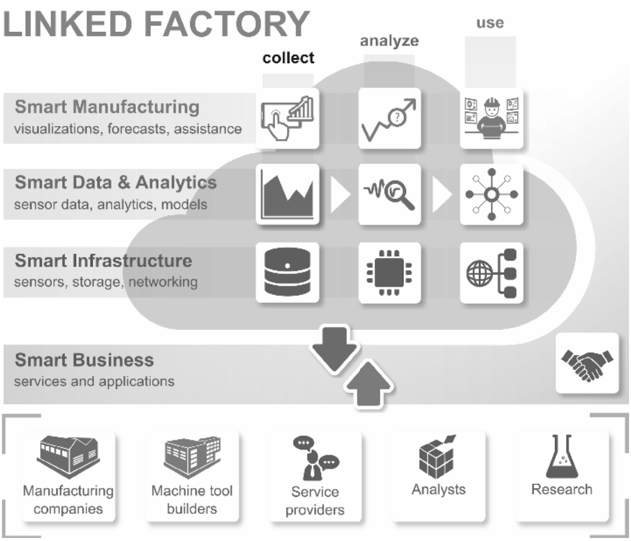


Fig. 12.7 LinkedFactory – Fraunhofer IWU holistic research approach for the factory of the future (Fraunhofer IWU)

- What new, innovative, technological and organizational solutions are required to implement consistent digitization of production?
- What data/information concepts need to be considered in order to provide the necessary transparency in production as a basis for decision-making?
- What technical as well as organizational barriers need to be overcome in order to implement a transferable overall system?

The realization of a robust overall system was achieved by means of a holistic approach where the modules supplement one another synergistically in the LinkedFactory (cf. Fig. 12.7), and can be combined with one another as tasks require. The goal is the real-time synchronization of material and information flow to allow for agile planning and efficient forecasting.

The following core areas were addressed:
- Possibility of a formal, standardized description of interfaces and operating languages/data sources in the production environment,
- Semantic annotation, model-based storage of data relevant for production,
- Task-related data analysis to provide information for production optimization,

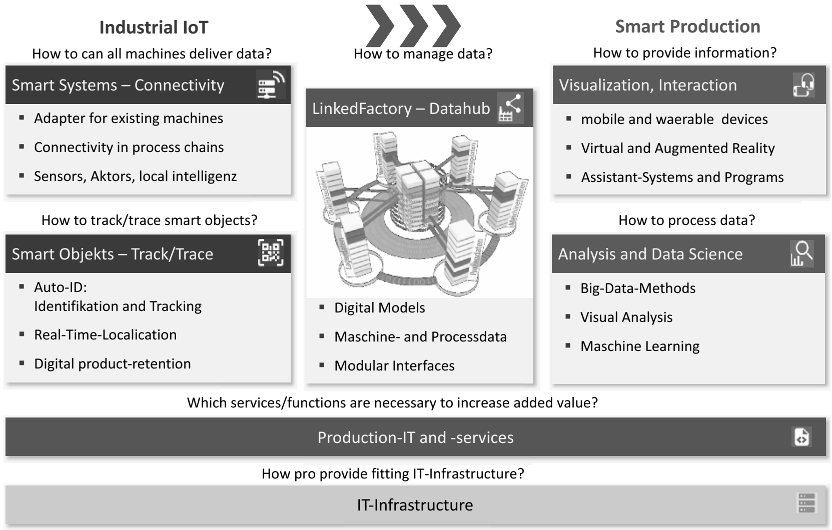


Fig. 12.8 Modular digitization toolbox – key questions and solution modules (Fraunhofer IWU)

- Innovative techniques for representing information, for formal interpretation, and for intuitive (bidirectional) human-machine interaction.

To realize the LinkedFactory concept on a task-oriented basis, Fraunhofer IWU developed a flexible modular toolbox (cf. Fig. 12.8). This enables the development of scalable connected systems in production. The modular toolbox here takes account of seven core questions identified in partnership with companies:

1. How can all machines provide data?
2. How is the data managed?
3. How is the data utilized/analyzed?
4. How is information made available?
5. How are smart objects localized?
6. What services/functions are required to increase value creation?
7. How can scalable IT systems be provided?

In terms of standardization, the modular digitization toolbox is guided by the Industry 4.0 reference architecture model (RAMI4.0) [16].

Selected sustainable solutions for the modular elements highlighted for production were developed within the SmARPro SmARt Assistance for Humans in Production Systems [17] research project sponsored by the BMBF.

In addition to flexible solutions for linking machines (Brown and Greenfield) as data sources (smart systems), the research focused on the SmARPro platform as a central data hub (entity of the LinkedFactory), SmARPro wearables as innovative solutions for visualization, and solutions for location identification.

Using CPS components, machines can be specifically incorporated into the extraction of data and information – sensors and communication elements allow the registration of data directly at the machine and in the process, enabling the transfer of the collected data to the LinkedFactory as a data hub. As shown in the transformation strategy depicted in Figure 12.3.1, particular attention is paid to being able to provide these to existing as well as new machines and plants as low-cost components. The goal is to facilitate pre-processing near the point of data origin. The solutions utilized in the context of standardized communication and data provision include both web technologies and solutions from the environment of machine-to-machine communications.

The focus of the overall LinkedFactory concept is the information hub as the central platform for data and services. It forms an integral element on the way to implementing innovative solutions to support flexible production structures. It integrates and links data beyond domains, for example with regard to

- Structure and design of existing machines/production systems and their relationships,
- Targets for controlling operation of the factory (PPS, ERP),
- Indicators and sensor information from current processes, and the processing status and outcomes of finished products (MES),
- Resource consumption of production components and of production and buildings infrastructure (control systems).

The aim is to link data managed specifically by domain and link it with other similar data in accordance with given specifications in order to infer new information or requested knowledge. This data is made available contextually via specified interfaces, depending on the role of the requester in question. By using the data thus linked and information generated on the basis of it, various services can be provided and combined. At Fraunhofer IWU, semantic web principles are being used for implementing these approaches in software [18]. One important property of this process is the formal representation of information using defined vocabularies, making it comprehensible to computers.

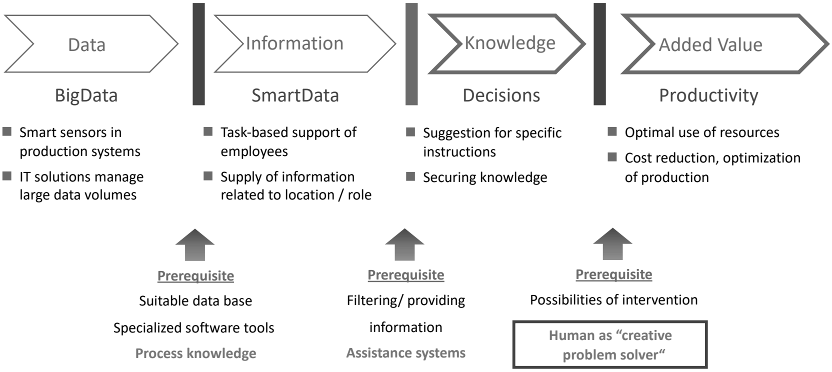


Fig. 12.9 Data as the basis for deducing information (Fraunhofer IWU)

Currently, the following support solutions, among others, are being implemented:

- Manufacturing a control that uses linked information from data sources within and outside of the company,
- Mobile contextual assistance systems for increasing product quality,
- Solutions for monitoring, controlling, and visualizing production processes,
- Mobile solutions for providing support in terms of servicing and maintenance.

A key aspect is that the data that is now becoming increasingly available may contain interrelationships that that have previously been unknown – they especially represent a large part of a company's technical manufacturing knowledge. These "hidden interrelationships" and the manufacturing-related knowledge, together with employees' experiential knowledge, represent an important basis for decision-making and planning processes (cf. Fig 12.9).

Further developments in information and communications technologies and new methods of data analysis, for example in the context of data mining or machine learning, may help to unearth "treasures" in the data stock generated within production-related IT systems during plant operation. Then the potential can be exploited for production-related savings or improvements [19].

The data flow entering the LinkedFactory need to be processed in a number of different ways in order to infer relevant information for creating added-value. The use of linked data technologies proves to be of great benefit since these technologies allow the connection of data flows from a range of resources – which are themselves occasionally subject to change over time. The basis for data processing here is

Station 1

Teil ID 23912912ue9wnqwkdn
Werkstückträger ID 123o81249ednlsdmc
Schwellenwert 123
Startzeit 12:05 Uhr
Planzeit 12:05 Uhr
Fehleranzahl 5

Station 2

Teil ID 23912912ue9wnqwkdn
Werkstückträger ID 123o81249ednlsdmc
Schwellenwert 123
Startzeit 12:05 Uhr
Planzeit 12:05 Uhr

Fig. 12.10 Role-specific and person-specific visualizations (plant overview, job details, notifications, directions) [20] (TYP4 Photography + Design, Phillip Hiersemann, www.typ4.net)

formed by a complex event processing engine (CEP engine). Rules are to be set by staff who understands the processes, bearing in mind the operational requirements and the available data flow. A key requirement in this context is that rules can easily be changed, which makes it possible to respond flexibly to process changes.

Using mobile terminals, information is provided to staff members contextually and depending on their current position, for example in the form of an augmented reality image.. The goal is to provide information that is directly relevant to the object in question. Work guidelines and information relevant for production can be received by employees without their workflow being interrupted. This changes the way information is displayed fundamentally. Information appears precisely when and where the individual needs it, without having to actively request it. The research here is focused on a broad range of different devices from tablet computers of various sizes, smartphones, up to smart watches and data goggles. Fig. 12.10 shows example interfaces for the contextual provision of information for assembly facilities. Here staff is provided with precisely the information required to better carry out their individual tasks for creating value, In order to be able to realize the location-specific provision of information via mobile terminals, it is necessary to identify the location of these terminals and match them against the areas for which information is to be displayed (regions of interest). Potential technological implementations include using WLAN-based range measurements, AutoID technologies, complex image recognition solutions, or simple QR codes. Individual solutions can be distinguished by the effort involved in their installation or by location accuracy, for example. Within the framework of the LinkedFactory concept, a standardized interface was developed that abstracts data from the localization technology utilized. Using the underlying linked data technologies, positional information for monitored devices is transferred to the LinkedFactory, and related to other data captured on an application-specific basis.

12.4 Challenges for CPS design

Due to the aforementioned heterogeneity of the overall system (caused by the architecture of a CPS system extending across domains and instances), the development of cyber-physical systems cannot be carried out from the perspective of a single specialist discipline. It requires a perspective that sets the multidisciplinary system at the center. One such comprehensive perspective, which extends beyond individual specialist disciplines, is provided by systems engineering.

Scarcely any explicit research has yet been carried out into how these kinds of intelligent systems are successfully developed, balancing requirements of time, cost, and quality. At that, the increasing intelligence within and networking between systems in particular as well as the accompanying multidisciplinary nature of approaches required pose new challenges for product development. Cyber-physical systems thus do not necessarily have firm system boundaries. Their functionality changes over the course of the product lifecycle; it often depends on application scenarios which occur during product usage on an ad hoc basis, so developers can only anticipate and take responsibility for it in limited measure. The often cited autonomous driving provides an example here where both individual vehicles as well as entire convoys can operate as autonomous systems with differing functionalities. Strictly speaking, however, only the convoy is a cyber-physical system for which however, as a rule, no one is directly responsible. Critical analysis is clearly lacking here as to whether companies in Germany will not only be able to invent and produce intelligent systems, but also successfully develop them in future in the face of global competition [21].

12.4.1 Systems engineering as the key to success

Systems engineering (SE) seems to be the right approach to overcoming the challenges described. Systems engineering is understood as the general interdisciplinary school of technical systems development which takes account of all the different aspects. It places the multidisciplinary system at the center and encompasses all of the different development activities. SE thus claims to orchestrate the actors in the development of complex systems. It addresses the product (and associated services, where applicable), production system (and value-added networks where applicable), business model, project management, and the operational structure. Systems engineering is thus extremely multi-faceted [22].

One particular focus of SE is the general and interdisciplinary description of the system to be developed, resulting in a system model. This includes an external presentation (such as diagrams) and an internal computer representation in the form of a digital model (the so-called repository). While data only appears once in a repository, it can be used multiple times and with various interpretations in external representations in order to generate specific views of the system. Model-based systems engineering (MBSE) places a cross-disciplinary system model at the heart of development. In the process, it does not exclude the existence of other models of the system, in particular those that are specific to one discipline, but it incorporates these via appropriate interfaces. Various languages (e.g. SysML), methods (e.g. CONSENS, SysMod), and IT tools (e.g. Enterprise Architect) are available for producing the system model, and they can even be variously combined with one another. So far there has not been any unified and recognized methodology in terms of an established school of MBSE [23].

This form of digital specification is simultaneously the basis for the ongoing seamless virtualization of the project activities when the project management is also founded on the information that has been generated correspondingly. The ongoing development of peripheral devices such as AR glasses permits new forms of workplace design for developers and an associated change in working methods in favor of efficiency and job satisfaction.

12.4.2 Performance level and practical action required

In practice, systems engineering as a term is very common, even if there usually is just a basic understanding. Only a few experts or specific fields such as software development possess a deep understanding of the existing methods and tools. In small and mid-size enterprises in particular, this expertise is only available via specific individuals. Even in larger companies, there is generally no companywide awareness of SE [24]. There are differences in particular in the different sectors. In aerospace engineering, systems engineering is firmly established. The automotive industry has recently been increasing its efforts here. SE programs have developed among leading OEMs out of company initiatives in mechatronic product development and are now being pursued systematically. In machine and plant engineering, by contrast, few companies are active in this field. Often, the upfront investments required for successful product development are still dreaded. Nevertheless, in the end it is only a matter of time before machine and plant engineering, too, will have to engage with this issue.

In practice the current performance level of systems engineering demonstrates a gap between aspirations and actual execution. In order to close this gap, the level of performance must be raised. This applies not only to new approaches coming out of research, but also to the usability of existing methods and tools. The following recommendations for action provide orientation for answering the question what fundamentally needs to happen in research and application:

- Consider all relevant aspects of development:
  Tomorrow's successful product creation will be characterized by universal processes and limited tool and method failures. Systems engineering may form the basis for this, but up to now it is more a collection of individual methods and practices. What is required is a holistic development framework is required, takes all of the different aspects of development (e.g. security by design, resilience by design, and cost by design) into account, not only at an early stage but also in an integrated way across the entire product creation process.

- Internalize product generations thinking:
  Strategic product planning sets the course for successful innovation early on. As an example, forward-looking and constantly updated release planning is indispensable for developing successful product generations. The continuous development of product updates and the parallel development of several product generations require a rethinking of product development and having a closer connection to product planning.

- Accelerating model-based product development:
  MBSE lies at the heart of a consistent SE approach. This requires different specialist departments and even different companies to be able to share and then process specific development information in the form of models. In order to realize this, existing languages, methods, and tools will not necessarily have to be

somehow integrated into one single standard (which in any case would be highly unrealistic). Instead, a new kind of exchange format will be required similar to the STEP format in the field of CAD, to be specified by an industry-led consortium.

- Combine PDMs/PLMs and MBSE into an integrated system model:
  Universal system data management throughout the product lifecycle such as the one provided by existing PDM/PLM solutions cannot continue alongside future SE or MBSE structures within the IT architectures of companies. PDMs/PLMs and MBSE need to be thought through and drafted in an integrated and synergistic way right from the start. Without the multidisciplinary perspective of MBSE, it will not be possible to establish PLM in future, while successful MBSE solutions are worthless if the models cannot be managed within an effective system.

- The development structure needs to become more agile:
  Whereas agile and flexible methodologies such as scrum and evolutionary prototyping etc. are already widespread in software development, they are rarely utilized for developing technical systems. It would seem that the agile software development paradigm cannot easily be translated since, unlike pure software products, functional prototypes cannot be designed in short sprints. There is a lack of continuous and effective adaptation of agile software development.

- Professionalize competency development in the field of SE:
  Germany's successful and specialism-oriented training urgently needs to be supplemented with a generalist training track. To this end, universities must create the necessary conditions across faculties such that this problem does not continue to be outsourced to practice. In addition, professional development certification courses such as SE-ZERT® need to be continually developed further and disseminated.

- Above all of the aforementioned recommendations for action stands digitization. It is the source of a revolution not only in production but at least as much in development work. Whether using AR goggles, virtual design reviews, big data, or assistance systems for data analysis, decision-making can be better founded since the data on what is happening within the project is more transparent, more up to date, and of higher quality. Developers are able to collaborate with each other more efficiently and in a more distributed manner within the company and beyond. Digital technologies and concepts thus need to be developed as quickly as possible and then applied productively to the systems engineering of tomorrow so that they can take a leading role in the future field of cyber-physical systems [25].

12.5 Summary and development perspectives

Digitization approaches for production and logistics within existing factory structures are currently only effective in isolated cases, specifically with respect to increases in efficiency, quality, and flexibility. In order to overcome this deficit, robust overall systems need to be realized where material and information flow synchronization facilitates production system structures with greater agility and flexibility, while retaining existing levels of productivity. In this context, possible starting points include achieving greater agility by means of event-based production control, the involvement of employees in specific real-time decision-making, and the resulting increase in flexibility of production processes. This can only be achieved if modular, task-oriented, combinable solution modules are available for implementing suitable support systems in production and logistics such that complexity can be mastered by means of a step-by-step process.

Nevertheless, for all the euphoria surrounding digitization, it is clear that in future people will still remain the guarantors of success and the central component of industrial value creation. This requires, among other things, improving knowledge of how to implement Industry 4.0 concepts in SMEs, developing suitable qualification offerings for (further) training, and planning sustainable factory structures, organizational forms, and processes to make the most of the opportunities provided by digitization.

All of the elements in the value chain will in future be changed by digitization. In cyber-physical systems, a universal structural approach is available that in future will also become established in smart, connected products. The structural equivalence of products and production systems will open up completely new opportunities in the design of value creation processes. Cyber-physical systems are the core element, characterized by the networking of smart production systems and smart products across the entire lifecycle.

This symbiosis allows for a range of new technical options such as using a product's sensory capabilities

- During its manufacture for process monitoring and control,
- During its use cycle for property adaptation and for diagnosing and assessing its condition, or
- For creating a data basis for improving and optimizing the design process.

Functional improvement will only be achieved by passing on knowledge and experience. The production optimization control loop is not only closed via information from the processes and plant of the production itself – data from the product's utilization phase also feeds into production optimization, e.g. for quality assurance in

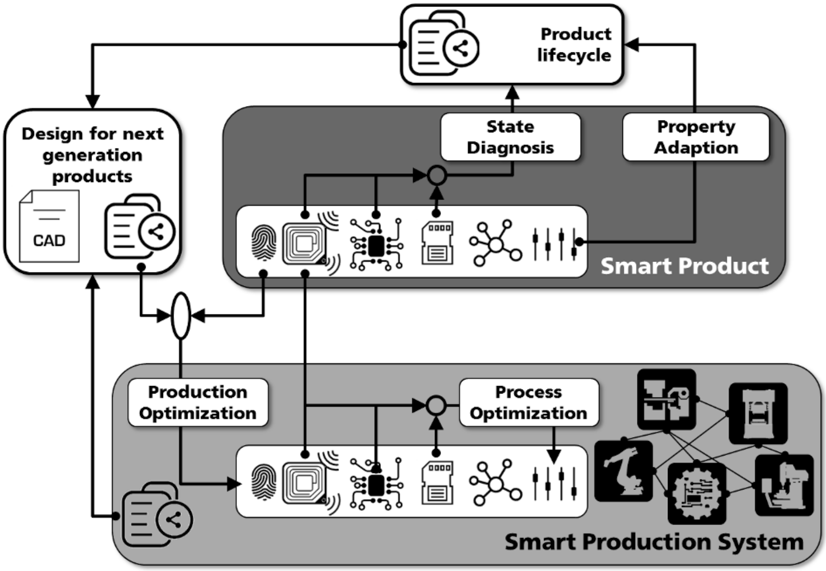


Fig. 12.11 Potential of the interaction between smart products and smart production systems (Fraunhofer IWU, Fraunhofer IEM)

the case of sensitive product properties. In the same way, data from the production process is combined with data from the product lifecycle to optimize product design, e.g. to facilitate the efficient manufacturing of the next product generation (cf. Fig. 12.11). These design and optimization processes are only efficient if they can be automated. This is where machine learning algorithms gain particular significance.

In this way, the evolution caused by Industry 4.0 technologies actually gives rise to a disruptive paradigm change. Production, equipment suppliers, and product developers enter a new quality of innovative cooperation. Production sites become co-developers of products, and equipment suppliers become the designers of the infrastructure for information and value streams.

Sources and literature

[1] Spitzencluster „Intelligente Technische Systeme Ostwestfalen Lippe (it's OWL)", BMBF/PTKA (2011-2017)

[2] ACATECH – DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN: Autonome Systeme – Chancen und Risiken für Wirtschaft, Wissenschaft und Gesellschaft. Zwischenbericht, Berlin (2016)

[3] Verbundprojekt „LiONS – System für die lichtbasierte Ortung und Navigation für autonome Systeme", BMBF/VDI/VDE-IT (2015-2018)

[4] Porter, M.E., Heppelmann, J.E.: How Smart, Connected Products Are Transforming Competition. Harvard Business Review (2014)

[5] Verbundprojekt „AcRoSS – Augmented Reality-basierte Produkt-Service-Systems, BMWi/DLR (2016-2018)

[6] Verbundprojekt „DigiKAM – Digitales Kollaborationsnetzwerk zur Erschließung von Additive Manufacturing, BMWi/DLR

[7] ACATECH – DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN: Smart Service Welt – Internetbasierte Dienste für die Wirtschaft. Abschlussbericht, Berlin (2015)

[8] Gill H. (2008) A Continuing Vision: Cyber-Physical Systems. Fourth Annual Carnegie Mellon Conference on the Electricity Industry Future Energy Systems: Efficiency, Security, Control.

[9] Roth A. (Eds.) (2016): Einführung und Umsetzung von Industrie 4.0 Grundlagen, Vorgehensmodell und Use Cases aus der Praxis, Springer Gabler Verlag

[10] Großmann K, Wiemer H (2010) Reproduzierbare Fertigung in innovativen Prozessketten. Besonderheiten innovativer Prozessketten und methodische Ansätze für ihre Beschreibung, Analyse und Führung (Teil 1), ZWF 10, S. 855–859.

[11] Großmann, K et al. (2010) Reproduzierbare Fertigung in innovativen Prozessketten. Konzeption eines Beschreibungs- und Analysetools (Teil 2), ZWF 11, S. 954-95.

[12] Hellmich A et al.: Drive Data Acquisition for Controller Internal Monitoring Functions, XXVII CIRP Sponsored Conference on Supervising and Diagnostics of Machining Systems, Karpacz, Poland (2016)

[13] Hammerstingl V, Reinhart G (2015): Unified Plug&Produce architecture for automatic integration of field devices in industrial environments. In: Proceedings of the IEEE International Conference on Industrial Technology S.1956–1963.

[14] Langer, T. (2015). Ermittlung der Produktivität verketteter Produktionssysteme unter Nutzung erweiterter Produktdaten. Dissertation Technische Universität Chemnitz.

[15] Grundnig, A., Meitinger, S. (2013): Führung ist nicht alles – aber ohne Führung ist alles nichts – Shopfloor-Manage¬ment bewirkt nachhaltige Effizienzsteigerung. ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb. 3, 133-136.

[16] Statusreport: Referenzarchitekturmodell Industrie 4.0 (RAMI4.0). Abgerufen von http://www.zvei.org/Downloads/Automation/Statusreport-Referenzmodelle-2015-v10.pdf

[17] SmARPro – SmARt Assistance for Humans in Production Systems – http://www.smar-pro.de

[18] W3C Semantic Web Activity, Online: http://www.w3.org/2001/sw/ [online 09/2015]

[19] Sauer, O. (2011): Informationstechnik in der Fabrik der Zukunft – Aktuelle Rahmenbedingungen, Stand der Technik und Forschungsbedarf. ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb. 12, 955-962.
[20] Stoldt, J., Friedemann, M., Langer, T., Putz, M. (2016). Ein Systemkonzept zur durchgängigen Datenintegration im Produktionsumfeld. VPP2016 – Vernetzt Planen und Produzieren. 165-174
[21] WIGEP: Positionspapier: Smart Engineering, Wissenschaftliche Gesellschaft für Produktentwicklung, 2017
[22] ACATECH – DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN: Smart Engineering. acatech DISKUSSION, Berlin (2012)
[23] Querschnittsprojekt Systems Engineering im Spitzencluster „Intelligente Technische Systeme OstwestfalenLippe (it's OWL)", BMBF/PTKA(2012-2017)
[24] Gausmeier, J,; Dumitrescu, R, Steffen, D, CZAJA, A, Wiederkehr, O, Tschirner, C: Systems Engineering in der industriellen Praxis. Heinz Nixdorf Institut; Fraunhofer-Institut für Produktionstechnologie IPT, Projektgruppe Entwurfstechnik Mechatronik; UNITY AG, Paderborn (2013)
[25] Verbundprojekt „IviPep – Instrumentarium zur Gestaltung individualisierter virtueller Produktentstehungsprozesse in der Industrie 4.0", BMBF/PTKA (2017-2020)

“Go Beyond 4.0” Lighthouse Project

13

Individualized mass production

Prof. Dr. Thomas Otto
Fraunhofer Institute for Electronic Nano Systems ENAS

Summary

Industry’s need for new technologies to provide differentiation and efficiency gains in production is the driving force of the Fraunhofer-Gesellschaft to pool competencies in order to provide technologies for success. The thus far rigid mass production will in future gain new impetus through digital manufacturing technologies such as inkjet printing and laser-based techniques, in particular. The integration of digital manufacturing technologies into a range of mass production environments will permit individualized production with zero setup times and only slightly increased cycle times.

Project overview

Aim of the “Go Beyond 4.0” lighthouse project

The aim of the project is to develop technologies for a resource-efficient and cost reducing machine-based individualization of series products within the advanced and connected mass production. This should enable the intelligent integration of digital manufacturing techniques in established and highly efficient process chains. This will facilitate small batch series through to batch sizes of one on a mass production basis, using a combination of industrial scalable and digital manufacturing techniques: digital printing as a material additive technique and laser machining as a thermal material removing technique. Thus, the productivity can be significantly increased compared with manual piece production. With the project’s success, the Fraunhofer-Gesellschaft is not only contributing to the continuation and further development of German competency in machinery and plant manufacturing but is additionally making a lasting contribution to the success of our national economy.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_13

Project partners

The following Fraunhofer institutes are involved in the project:

- Fraunhofer Institute for Electronic Nano Systems ENAS
- Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM
- Fraunhofer Institute for Laser Technology ILT
- Fraunhofer Institute for Applied Optics and Precision Engineering IOF
- Fraunhofer Institute for Silicate Research ISC
- Fraunhofer Institute for Machine Tools and Forming Technology IWU

Research schedule/sponsorship

€8 m. (+ €1 m. as required) via the Fraunhofer-Gesellschaft

Contact

Prof. Dr. Thomas Otto, Fraunhofer ENAS
Tel. +49 371 45001 100

13.1 Introduction

Across industries, the demand for innovative, individualized components for the future markets of production equipment, automotive and aerospace engineering, and lighting is continuously growing. Special functional materials are used to provide the corresponding components with the necessary highly qualified functionalities, with applications focused clearly on electronic and optical functionalities. The anticipated efficiency stemming from the use of high-quality organic and inorganic materials is the driving force for the development of customized process chains for the manufacture of intelligent components with high diversification.

This demand is a global phenomenon; currently, in highly developed industrialized nations, comprehensive economic structural measures are taken. Of particular

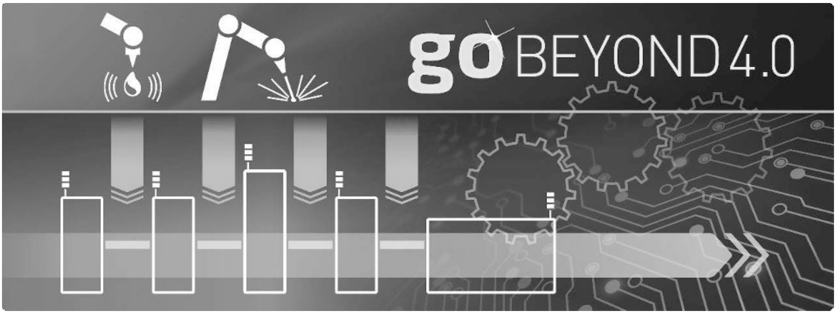


Fig. 13.1 Key image for “Go Beyond” (Fraunhofer ENAS)

note here are the revitalization of industrial manufacturing in the USA and the development of flexible manufacturing technologies in China.

The diversification of products requires new manufacturing strategies that must address the following challenges:

- Increase of product diversity, batch sizes shrink to batch size one (unique items)
- Products become intelligent (the capture, processing, and communication of data)
- The necessary component intelligence is produced by the material-efficient integration of new functional materials
- Environmentally-conscious recycling of corresponding products

These challenges are met and overcome in the Fraunhofer “Go Beyond 4.0” lighthouse project by integrating digital manufacturing techniques into existing mass production environments.

The IT requirements for this novel kind of integration will result from the holistic networking of production in the course of the development of the industrial Internet (Industry 4.0).

On that basis, a new process chain design can be established which allows for an overlap of industries and products – made possible through the combination of first-rate competencies from the Fraunhofer associations for production (IWU), materials and components (IFAM, ISC), microelectronics (ENAS), and light and surfaces (IOF, ILT). This solution approach directly addresses industry’s requirement for efficient manufacturing processes, facilitating batch size one fabrication in the context of highly efficient mass production strategies.

13.2 Mass production

Mass production facilities are designed to produce the largest possible numbers of standardized components and products with the lowest cycle times. In order to achieve this, tool-based manufacturing technologies are lined up in assembly lines. The scope and complexity of the assembly line vary depending on the complexity of the specific component or product. In the case of changeover to another product, time-consuming machine resetting is required, which in turn reduces production efficiency.

Market demands for individualized parts and products inevitably lead to smaller volumes (down to batch size one), which in turn necessitate more frequent changes of process chains. This additional effort significantly reduces the efficiency of mass production. Furthermore, it is to be expected that in future constant changes in the products to be manufactured will shape daily business.

Advanced mass production in the traditional sense thus requires rethinking. Manufacturing strategies need to be developed that maintain the economic benefits of mass production, while simultaneously permitting significant reductions in batch sizes down to extremely short runs and even one-of-a-kind products. A relevant approach in the Fraunhofer-Gesellschaft “Go Beyond 4.0” lighthouse project is based on the methodology of digital manufacturing techniques and integrating these into the optimized mass production environments of the rapidly developing industrial Internet.

13.3 Digital manufacturing techniques

Prof. Dr. Reinhard R. Baumann · Dr. Ralf Zichner
Fraunhofer Institute for Electronic Nano Systems ENAS

Today, product development is extensively mapped in computer systems. By combining design and construction data, complete computer models of new products are generated. These product datasets are fed into output systems that either visualize or transfer them into real, physical objects. Data visualization already takes place during the initial development steps realized with the aid of advanced monitors, which are tailored to the particular requirements of product display. In order to objectify products that have hitherto only existed virtually, technologies need to be selected that result in real components.

Traditionally, “aids” are produced from the data, which then allow real components to be developed from raw materials via the standard manufacturing techniques (primary forming, shaping, separation, joining …). These “aids” are often molds and auxiliary constructions that are labor-intensive to produce; the financial outlay to develop them is compensated for by allocation across large batches.

Usually, if digital manufacturing technologies are chosen for objectification, the geometries of the manufactured component can only be changed if the dataset transmitted to the manufacturing system is adapted. Modern CNC (Computerized Numerical Control) machine tools fulfill this requirement by choosing suitable tools for material removing via machining. This results in large quantities of material to be recycled.

In another group of digital manufacturing technologies, material is only applied to those points where it is needed for building up the component geometry; Think, for example, of building up the component geometry via laser sintering of metal powder. Included in the group of additive manufacturing methods are also digital printing technologies. In the case of inkjet technology, suspensions of nanoparticles in the form of droplets (with volumes in the picoliter range) are layered on top of one another. After evaporation of the carrier fluid and a sintering phase (often photonic), stable three-dimensional component geometries are produced with specific functional properties such as electrical conductivity. This allows, for example, the manufacturing of free-form conductor tracks on complex component structures.

The previously described printing and laser-based techniques allow especially in the component manufacturing with high-performance materials for yet unknown levels of material use efficiency while maintaining one-of-a-kind component geometries.

13.3.1 Digital printing techniques

Prof. Dr. Reinhard R. Baumann
Fraunhofer Institute for Electronic Nano Systems ENAS

In the 500 years since the invention of the printing press by Johannes Gutenberg, generations of technicians have developed the picture-by-picture transfer of ink onto substrate to such a technical degree that today the human eye generally perceives printed images as halftone objects just like the natural original. Despite the fact that printed products consist of a high-definition cloud of microscopic halftone dots. Traditional printing processes (letterpress, gravure print, screen printing, and lithography) use hard-wearing printing blocks that are equipped with the image or text content for reproduction one-time. From there identical copies are mass produced in high numbers during the actual printing process. Production of the printing blocks is technically elaborate and only becomes economically beneficial if this work can be allocated across as large a number of copies as possible.

In order to avoid this effort, two alternative pathways have been pursued in the last 100 years towards producing images on a substrate that the human eye perceives as a visual whole. In both techniques, the picture elements are generated each time a copy is produced. Printers refer to this production strategy as "single edition printing" while production technicians rather refer to it as "batch size one". This describes the fundamentals of digital printing.

This promising approach is based on the idea of producing picture elements (halftone dots) by using small droplets of colored liquid. Nevertheless, the real challenge of this printing process, nowadays known as inkjet, is the precise pictorial placement of the droplets on the substrate and the reproducible production within narrow boundaries.

While in traditional printing processes the printing block is imaged once and subsequently produces many copies (image one – print many), in the digital Inkjet printing process printing blocks are no longer required; the ink image is transferred once onto the substrate (image one – print one).

This results in a unique geometric distribution of ink on the substrate every time the printing process is run. It is crucial, that for each individual illustration cycle a new unique dataset can be taken as the basis. Exactly this is facilitated by modern digital data processing systems, which allow for cycle times within the digital printing process with substrate speeds of up to several meters per second. These digital data systems were the force behind the conceptualization of *digital printing*.

Digital printing processes essentially fulfill the technological requirements for low-volume manufacturing down to batch size one. So far, the geometrical requirements of the substrates have addressed the human sense of sight (i.e. the human eye) with their functionality of color, now they must be able to address additional functionalities such as electrical conductivity or insulation. These functionalities are also addressed by producing an "image", but instead of a landscape, the image may be a conductor track capable of carrying electric current, for example. The discovery that digital printing can produce material patterns with different functionalities at high rates of productivity gave rise to functional printing over the last 25 years. The approach is pursued by equipping ink systems with new properties such as electrical conductivity in order to form systems of layers, according to the principles of printing, which then can be used as electronic components (resistors, capacitors, diodes, transistors, and sensors) or simple circuits. Another central challenge is, aside from the corresponding adaptation of the printing processes, the preparation of the inks. Today, as a result of the multifaceted development work conducted, functional printers are able to choose between two kinds of functional inks that are commercially available.

The first kind is suspensions of nanoparticles, which determine the functionality of the ink. The second kind is solutions of functional molecules. Commercially available inks permit nowadays, among other things, the manufacture of corresponding material patterns made of conductive and semi-conductive organic polymers.

Today, the field of digital functional printing technology is dominated by inkjet systems. Due to the ink formulation benefits the so-called Drop-On-Demand (DoD) inkjet is preferentially used; there again, printing systems based on piezoelectric actuators (MEMS technology) dominate, since they avoid extreme thermal loads on the inks compared to bubble jet printing.

Inkjet technology has experienced an enormous boost from the graphical industry in the last decade. Today, all notable manufacturers of printers have inkjet printing systems in their portfolios or already integrated them into their traditional printers. All optimizations of the inkjet process will aid to establish and further develop digital functional printing, permitting an economical manufacturing of batch size one in mass production environments.

13.3.2 Laser-based techniques

Dr. Christian Vedder
Fraunhofer Institute for Laser Technology ILT

No other tool can be metered and controlled nearly as precisely as the tool of light. Currently, lasers are used in a wide range of fields ranging from telecommunications and measurement engineering, to production of electronic microchips to ships. Alongside classical laser-based techniques such as cutting, drilling, marking, and welding, laser technology has also facilitated new production processes. These include selective laser beam machining for prototype construction, laser structuring and laser polishing of component surfaces, laser deposition welding e.g. in turbo engine production, selective laser etching, and EUV lithography; leading to the opening of new markets. Exemplary innovative applications are the laser-based generative manufacture of functionally- and resource-optimized metal components using "3D printers" and the extensive direct micro structuring of functional surfaces via high-performance short-pulse lasers (photovoltaic, OLED, friction- and wear-optimized surfaces).

In contrast to conventional processes, lasers allow the cost-effective manufacture of both small quantities as well as complex products. The to a large extent existing independence of production costs from quantity, wide variety, and product complexity offers huge economic advantages. This enables high-wage nations such as Germany to remain globally competitive by means of innovative products with demand-optimized properties such as functional integration. For this purpose, laser sources and processes in the field of surface technology have been the subject of continuous research to accelerate the progress of industrial implementation of individual, additive functional integration in components, realized by combining digital printing techniques and laser-based pre- and post-processing techniques.

In laser beam processing, the optical energy emitted by the laser is converted into thermal energy by absorption in the work piece (in this case, component surface or printed functional layers on a component). Alongside many other dependencies, this absorption is primarily dependent on the wavelength and material. Thus, not all laser types are equally suitable for e.g. surface structuring via laser ablation or

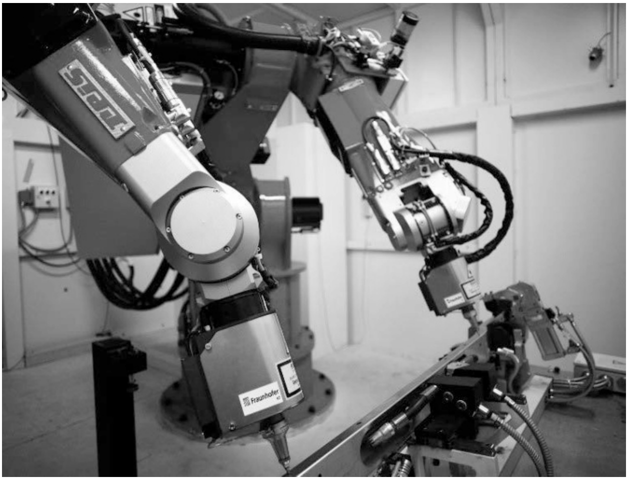


Fig. 13.2 Two robot-operated laser systems (Fraunhofer ILT, Aachen)

thermal functionalization of printed functional layers, as this depends on their capability to emit ultraviolet, visual and up to infrared light continually or in pulses.

Just as with digital printing techniques, component-individual material processing is possible by using digital laser-based techniques. Furthermore, lasers represent a contactless and pressureless tool with performances in the micro- to kilowatt range while only wearing down to a minimum extent. Laser also offers high kinematic flexibility, making it suitable for automation and in-line system integration into existing mass production environments.

Surface structuring via laser ablation is applied in this project in order to purposefully open up the surfaces of metallic and polymer mass-produced components and thus enable the embedding of the later printed functional layers. Additionally, the targeted removal of material supports the functionality of hybrid-integrated components such as piezoelectric actuators as well as the preparation of the component's surface by cleaning, improving the wetting properties, and mechanical or chemical bonding conditions etc. for subsequent printing. For this purpose, short- to ultrashort-pulse solid-state lasers are utilized with pulse lengths ranging from several nano- to femtoseconds. High levels of beam quality and beam power densities make it possible to realize structural sizes of several micrometers up to several nanometers at high process speeds and precision. Beam division by means of diffractive optical elements as well as systematic direction of separate or combined partial beams facilitate the parallel processing of identical 3D components and also the individualized structuring of varying 3D components. Within the framework of this project, aluminum components as well as carbon/glass fiber-reinforced plastics (CFRPs/GFRPs) are structured and cleaned where applicable.

The layers applied to the component via wet chemical digital printing require subsequent thermal treatment, in which liquid constituents (e.g. solvents) and binders used to stabilize the for the printing process necessary inks or pastes (e.g. organic vehicles that prevent the agglomeration or sedimentation of the functional particles in the solvents) evaporate.

This procedure further allows sintering or fusing of the functional elements such as silver particles. The printed layers gain their desired functionality, such as electrical conductivity, only as a result of this subsequent thermal treatment.

In addition to traditional thermal treatment processes with tools such as furnaces, infrared heaters, or flashbulbs that permit rapid, large-surface treatment of printed layers, laser treatment has the benefit of temporally and spatially selective subsequent thermal treatment. Thus, with the appropriate choice of laser source, the temperatures required for functionalization of the layers – temperatures which often lie far above the thermal damage threshold of the temperature-sensitive substrate underneath or the hybrid integrated electronic components within close proximity – can be achieved temporarily, without permanently damaging the latter. In the framework of the project, printed electrical insulator, conductor, and sensor structures as well as optical reflector layers on aluminum, CFRP/GFRP and ORMOCER components are thermally treated using laser radiation.

13.4 Demonstrators

13.4.1 Smart Door

André Bucht
Fraunhofer Institute for Machine Tools and Forming Technology IWU

In the automotive industry, the market-based trend for individual products encounters rigid manufacturing chains. Production is strongly tool-related in order to guarantee efficient manufacturing of large batch sizes, but offers only limited opportunities for individualization. The functional individualization is mainly achieved by installing mechatronic systems such as actuators, sensors, and control units. First, this differential construction method leads to significantly increasing assembly effort as, for example, the installation of the cable tree is nowadays one of the most exten-

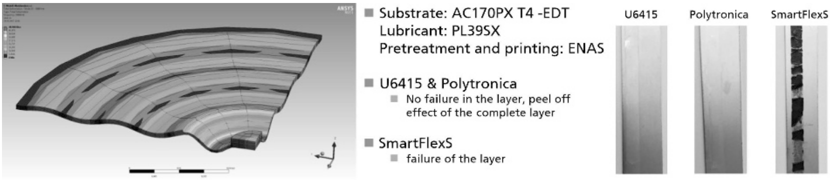


Fig. 13.3 Finite element analysis of the printed ultrasound transducer (left); printed insulation layers and conductor tracks after forming (right) (Fraunhofer IWU)

sive steps in vehicle assembly. Secondly, adding further flexibility to the already existing manufacturing structure results in exponentially increasing complexity costs in the fields of logistics, development, and production. Thirdly, thinking in terms of single components leads to increased installation space and weight requirements.

Thus, there are limits set on the desire for increased individualization and functional integration.

As mentioned in previous paragraphs, digital manufacturing steps offer a potential solution. The integration of digital process steps into analog tool-based manufacturing chains allow the individualization of batch sizes down to one.

With the aid of the technology demonstrator SmartDoor, these possibilities are developed and demonstrated. The design and functionality of this demonstrator is based on a real vehicle door. Functional elements such as ultrasound transducers, conductor tracks, and controls are applied using printing measures on both, the exterior component manufactured via forming technology as well as the fairing component manufactured via injection molding. The functional elements are manufactured in a hybrid process chain as a combination of analog and digital process steps.

The design of the functional elements to be printed is based on actual requirements of a vehicle door. For this purpose, typical industry-related requirements as well as the necessary physical parameters were taken as a basis. It could be demonstrated, that hitherto mechatronic sensors and actuators such as ultrasound transducers are basically implementable as printed layer construction. But to achieve the performance level of currently used systems, further optimizations with regard to the design of the system and the properties of the printed layers are necessary. A comparison to the current state of the art pointed out, that increased resilience and improved physical properties of the printed layers are necessary. Furthermore, the integration in analog process chains – in this case forming technology –requires a significant leap in the productivity of digital processes. In the further course of the project, the focus will mainly be put on these aspects.

13.4.2 Smart Wing

Dr. Volker Zöllmer
Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM

Carbon fiber- and glass fiber-reinforced plastics (CFRPs and GFRPs) are characterized by their high specific stiffness and strength while simultaneously maintaining low weight. Thus, their application in lightweight structures increased. However, the advantages of these lightweight materials cannot yet be fully exploited: on the one hand, quality fluctuations occur due to partially manual process chains; on the other hand, components made of fiber-reinforced plastics (FRPs) cannot be analyzed using the usual nondestructive testing methods. As in contrast to metallic structures, for example, damages caused by impacts during operation cannot be clearly recognized or detected.

For this reason, structural components made of FRP are construed with large safety margins or require short service intervals while in use to guarantee sufficient reliability. Both lead to increased costs. Structural health monitoring (SHM) during operation – e.g. by equipping fiber-reinforced structures with sensors- can reduce safety factors and thus costs and energy use during operation. This means that damage can be identified not only at the surface but ideally also inside the component, provided that the stability will not be compromised by the FRP’s construction. The integration of very thin film-based sensors, however, is already problematic since in extreme cases these can lead to delamination within the FRP component and thus to the breakdown of an FRP structure.

Fig. 13.4 View of Fraunhofer IFAM’s assembly line for the digital functionalization of FRP components (Fraunhofer IFAM)

In the course of the subproject B “Smart Wing”, digital printing processes are used to apply sensors on components to monitor the occurring load stresses, and further, to integrate these sensors at relevant points during manufacture inside the FRP components. Integrated sensors offer the possibility to monitor the fiber-reinforced components permanently, detecting high loads and damages early. Additionally, sensors to detect icing are integrated as well as heating structures to remove them. Digital printing and laser processes permit to print and functionalize sensor structures directly and locally *onto* FRP surfaces. An integration of electrical, sensor, or capacitive functions right *into* the fiber composite is also possible: structures made of functional materials can be applied with high resolution directly onto non-wovens or fabric via digital printing processes and further be used as an impregnable textile layer in the manufacturing process of the fiber-reinforced composite. This way, in the vacuum infusion process printed polyester and glass fiber non-wovens are processed into functionally integrated GFRP. Carbon fibers, however, need to be electrically insulated first. For this step, however, printing processes are also suitable, enabling insulating and barrier materials to be applied directly onto the fibers. In general, the materials applied during the printing process require subsequent thermal treatment. A subsequent local thermal treatment of the printed structures using lasers or high-energy UV radiation is suitable.

13.4.3 Smart Luminaire

Dr. Erik Beckert
Fraunhofer Institute for Applied Optics and Precision Engineering IOF

The demand for custom LED lighting systems is constantly growing in, among others, the automotive, medical technology, industrial manufacturing, and interior and exterior architectural sectors. This demand requires optical components and systems that, according to the application, illuminate specific areas and generate within these areas defined, individual lighting patterns. These lighting patterns can serve as information and interaction purpose, but also support the wellbeing of the user.

In subproject C “Smart Luminaire”, this demand is addressed. Based on standard optics, individual optical components manufactured via laser and digital printing processes are investigated. Using the inkjet printing process, the optical hybrid polymer ORMOCER® is applied in layers onto a standard optical system or any other substrate and subsequent hardened by UV lighting or infrared laser exposure, respectively. This results in three-dimensional refractive optical structures with

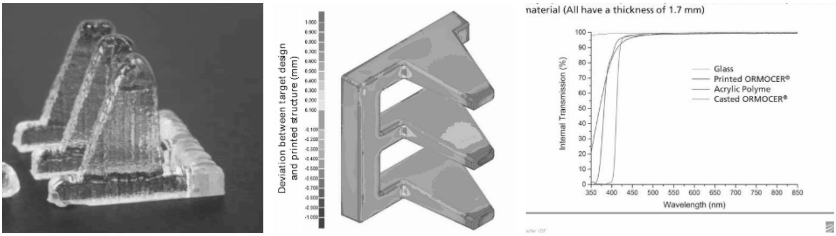


Fig. 13.5 Prototypes of 3D-printed optics (left), measurement of geometrical deviation using computed tomography (center), comparison of transparency (right) (Fraunhofer IOF)

dimensions in the millimeter or centimeter range that, with optimum process parameters, are comparably transparent to optical bulk materials made of glass or polymer. The transparency of the printed optical structures, as well as the required shape accuracy and surface roughness represent particular challenges in the process design.

These three-dimensional refractive optical structures are combined with diffractive structures, which are printed on the surface of the inkjet-printed optic using two-photon absorption. Furthermore, printed electrical conductor tracks are integrated into the optical structure to allow the integration of hybrid optoelectronic components such as LEDs or photodiodes via precision assembly and contacting. These are partially or completely embed into the optical structure in the subsequent printing process. LEDs and photodiodes permit the interaction of the lighting components with their environment. This takes place via visual and sensor functions that display system states or measure environmental parameters. The approach to manufacture the optical component digitally thus not only addresses their application-oriented individualization but also the optoelectronic system integration.

This allows the manufacture of completely new, individual, and highly integrated components and LED-based systems for structured lighting without reconfiguring times of the machine during production. The possibility to address individual customer requests opens new application fields for modern LED light sources and supports the further distribution into fields of consumer and industrial lighting technology.

13.5 Summary and outlook

The in the framework of the lighthouse project acquired technological progress – to efficiently link individualized production with the economic benefits of mass production – opens up new and promising perspectives for the production sites Germany and Europe.

The flexibility of the digital manufacturing processes used (printing and laser-based techniques) permit the manufacture of component geometries of practically any shape – from small lighting objects (Smart Luminaire), to macroscopic objects (Smart Door) up to larger objects (Smart Wing). Further do these techniques permit the integration of microelectronic components (microcontrollers, data stores, communications units, etc.) in or on the objects to be manufactured. By using these hybrid technologies, objects/products/components can gain additional functionalities that guarantee system intelligence.

The integration of printed structures, functions, and even microelectronics into components requires a particular degree of component reliability. Within this project, this topic is analyzed and evaluated in order to derive guidelines for the technological steps of functional printing, laser-based techniques, integration of microelectronics, and complete digital automation.

Modularity, integrability, and reliability of the digital manufacturing processes will contribute to an even more effective utilization of machinery in the future. The technologies developed by Fraunhofer are suitable to be integrated flexibly into existing manufacturing lines according to the modular principle. Thus, already existing manufacturing lines can be improved with regard to efficiency and capacity.

Exploitation plan

The exploitation of the project's results is based on three models:

- Direct exploitation of the results in industry via research services, technology transfer, and licenses
- Exploitation via an online technology-atlas provided to industry
- Exploitation via a Fraunhofer-led Application Center

The direct exploitation of the results in industry started simultaneously to the project work. Here, individual Fraunhofer technologies are transferred, for example, to efficiently manufacture components for aerospace and automotive engineering and equip them with new functional properties. Aim is to establish Fraunhofer as a technology brand for individualized mass production over the course of the following years. The technology exploitation will take place by means of suppliers and original equipment manufacturers (OEMs).

Exploitation of the project results via an online technology-atlas provides an opportunity for various manufacturers to determine, assess, and combine modular Fraunhofer technologies in a way that the demand for new, intelligent, and completely digital production lines can be met. Role model for the visualization of the technology-atlas is the Fraunhofer Shell Model[1].

Exploitation of the project results via Fraunhofer-led application centers provide a demonstration to industry about the capability of Fraunhofer technologies in real industrial environment. The intention is to transfer Fraunhofer technologies over a period of a few years to industry following the project’s completion.

---

[1] Weblink to Fraunhofer Shell Model: https://www.academy.fraunhofer.de/de/corporate-learning/industrie40/fraunhofer-schalenmodell-industrie-4-0/_jcr_content/contentPar/sectioncomponent/sectionParsys/textwithasset/imageComponent/image.img.large.png/1500464510890_Industrie-Layer.png

Cognitive Systems and Robotics

14

Intelligent data utilization for autonomous systems

Prof. Dr. Christian Bauckhage
Fraunhofer Institute for Intelligent Analysis and
Information Systems IAIS
Prof. Dr. Thomas Bauernhansl
Fraunhofer Institute for Manufacturing Engineering and
Automation IPA
Prof. Dr. Jürgen Beyerer
Fraunhofer Institute of Optronics, System Technologies,
and Image Exploitation IOSB
Prof. Dr. Jochen Garcke
Fraunhofer Institute for Algorithms and
Scientific Computing SCAI

Summary

Cognitive systems are able to monitor and analyze complex processes, which also provides them with the ability to make the right decisions in unplanned or unfamiliar situations. Fraunhofer experts are employing machine learning techniques to harness new cognitive functions for robots and automation solutions. To do this, they are equipping systems with technologies that are inspired by human abilities, or imitate and optimize them. This report describes these technologies, illustrates current example applications, and lays out scenarios for future areas of application.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_14

14.1 Introduction

Monitoring complex processes, analyzing them intelligently, and enabling them to make the correct decisions independently even in unplanned or unfamiliar situations: this is the goal that Fraunhofer experts are currently pursuing with the aid of new cognitive functions for robots and automation solutions that utilize machine learning (ML) methods. The basic concept is for systems to be equipped with technologies that are inspired by human abilities or imitate and optimize them. ML methods demonstrate their full added value where the parameters of a process are not (or not fully) known, where these frequently change, and where the complexity of a process is so great that it can neither be modeled nor implemented as a fixed process. Based on sensor data analyzed in near-real time, ML enables systems to constantly adapt the process and continuously improve its performance through on-going learning.

When we humans see an object, for example, we can use criteria from our learned experiences such as shape, size, color, or more complex characteristics to confirm whether it really is the intended object, even if we have not seen that specific peculiarity before. To do this, humans draw on an accumulated wealth of experience.

Learning systems are already benefitting from technologies similar to this human behavior in numerous sectors. The basis for this are large volumes of data that are processed with the aid of various methods, analyzed in near real time, and utilized for various application scenarios. Thanks to significant increases in processor capacities, data analysis that far exceeds human capabilities is now possible, facilitating the identification of extensive relationships and patterns. Using this knowledge, processes in production and automation as well as in the service sector or home environment can be optimized for user requirements and executed with a very high degree of autonomy.

14.2 Fundamental and future technologies for cognitive systems

Prof. Dr. Christian Bauckhage
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

Recently, we have seen rapid progress in the field of machine learning that has led to breakthroughs in artificial intelligence (AI). Above all, this development is driven by deep neural networks. Neural networks are complex mathematical decision-making models featuring millions of parameters that are optimized during a training phase. To do this, statistical learning techniques, very large training datasets (e.g. sensor, text, or image data), and powerful computers are utilized. Once this has been done, these neural networks are in the position to solve cognitively demanding problems [1]. In image analysis [2], speech recognition [3], text comprehension [4], or robotics [5], a level of performance is now possible that approaches or even exceeds that of the human brain (e.g. in medical diagnostics [6] or in gaming scenarios [7][8]). The state of the technology can thus be summarized briefly and succinctly using the following formula:

Big data + high performance computing + deep architecture = progress in AI

In order to understand why this equation works out and what future developments it leads us to anticipate, we want to answer the following questions here: what are artificial neural networks? How do they work? Why have they suddenly become so good? What should we expect from this field in future?

14.2.1 What are artificial neural networks?

Artificial neural networks are mathematical models that can be implemented on computers and carry out a form of information processing that resembles the human brain. Put simply, these models consist of numerous small processing units or neurons that are networked together. An entire network of neurons combines to form a complex processing unit capable of classifying data or making forecasts, for example.

The schematic illustration in Fig. 14.1 demonstrates how each individual neuron of a neural network performs a comparatively simple mathematical function that maps input values to an output. The precise values of the input and output obviously depend on the application in question. However, since data (sensor measurements, text, images, etc.) is always represented within a computer's memory by numbers, artificial neurons are designed to process and output numbers.

The input numbers are first multiplied by weight parameters and added up. The result of this synaptic summation then undergoes a nonlinear activation function in order to calculate the output. A classic example of this kind of activation function can also be seen in Fig. 14.1. The sigmoid (S-shaped) activation function shown here results in the neuron's output being a number between -1 and 1. We can thus imagine that a single neuron is making a simple yes/no decision: if the weighted sum of the input values is larger than a threshold value, then the neuron produces a

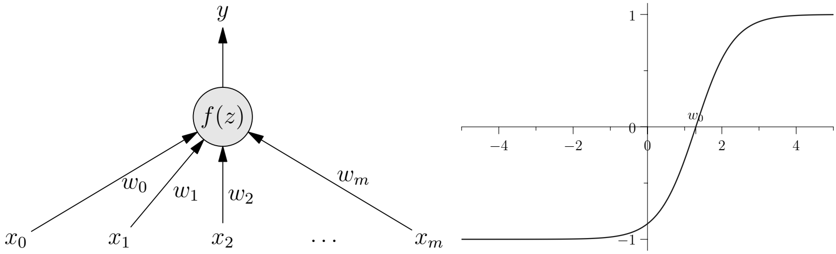


Fig. 14.1 Schematic representation of a mathematical neuron (left), and example of an activation function (right) (Fraunhofer IAIS)

positive number; otherwise it produces a negative one. The closer this output is to either of the extreme values 1 or -1, the more certain the neuron is that its weighted input lies above or below the threshold value.

Individual neurons are then typically arranged in layers and connected with one another. This produces neural networks such as that shown in Fig. 14.2 which we can imagine as essentially calculating a very large number of parallel or consecutive yes/no decisions. If neural networks are large enough and their weight parameters are set appropriately then they can solve practically any problem imaginable in this way.

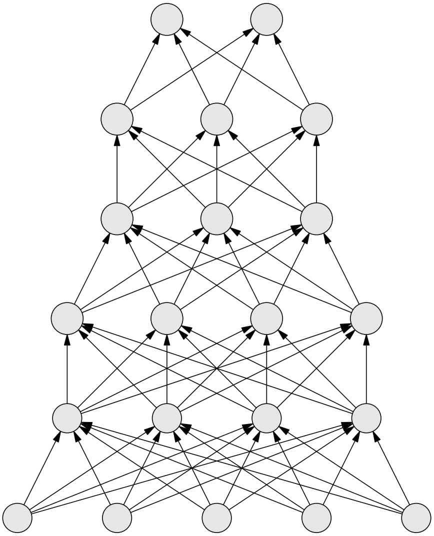


Fig. 14.2 Schematic representation of a hierarchical neural network; circles represent neurons and arrows symbolize synapses, i.e. connections between neurons. Information processing in this kind of network always takes place in the direction of the arrows, in this case from left to right. The first layer of this network receives the input, performs calculations on it, and passes the results on to the next layer. The final layer produces the output. Modern deep neural networks consist of hundreds of layers with hundreds of thousands of neurons each and can thus carry out complex calculations and process complex data. (Fraunhofer IAIS).

As this was already shown mathematically in the 1980s [9][10], the question of course arises as to why neural networks have only recently begun to be used universally with great success. In order to answer this question, we first need to consider how a neural network learns to solve a desired task and what learning means in this context.

The best way to do this is to take a look at a simple, concrete example. Let us assume that a neural network like the one in Fig. 14.2 is designed to recognize whether a dog or a cat is shown in an image with a resolution of 256x256 pixels. To do this, the network's input layer must first and foremost be significantly larger than that shown above, containing a minimum of 256^2 neurons. It would also certainly have to consist of more than six layers in order to be able to solve this apparently simple but actually very demanding problem. The output layer however could still consist of two neurons as in Fig. 14.2 because we have the freedom as developers to specify which subject should produce which output. The following output coding, for example, would make sense here: [1,-1]for images of dogs, [-1,1] for images of cats, and [0,0] for images containing neither dogs nor cats.

In order for the neural network to adequately solve this classification problem, the weight parameters of its neurons must be set according to the task. To achieve this, the network is trained using examples. The training requires a dataset that contains pairs of possible inputs and corresponding desired outputs, in this case images of dogs and cats together with the corresponding output codes. If this kind of training data is available, the training can be carried out using an algorithm that proceeds as follows: starting with randomly initialized weights, the network calculates an output for each training input. At the start of the training, this typically shows significant deviation from the desired output. Using the difference between the calculated and target outputs, however, the training algorithm is able to calculate how the network's weights need to be set so that the average error is as low as possible [11]. The weights are thus automatically adjusted accordingly, and the next round of training begun. This process is repeated until the network has learned to produce the desired outputs.

In practice, this training methodology has for a long time been faced with two fundamental problems. On the one hand, the calculations that are required to adjust the weights of a neural network are extremely elaborate. In particular large or deep neural networks with numerous layers of neurons could thus not be trained within reasonable timescales on earlier generations of computers. On the other hand, one of the basic theorems of statistics states that statistical learning processes only function robustly if the number of training instances is significantly greater than the number of parameters of the model [12]. To train a neural network with 1 million weight parameters, for example, you would need at least 100 million training examples so the network can learn its task correctly.

In the era of big data and powerful, inexpensive computers/cloud solutions, however, both problems have been solved such that this cognitive technology can now realize its full potential. Indeed, neural networks can be employed universally and are able to solve classification, forecasting, and decision-making problems that are far more complex than our simple example.

14.2.2 Future developments

Alongside the simple so-called feed-forward networks that we discussed above, there is a whole range of further varieties of neural networks with the potential to spur on additional technological developments and open up new areas of application. It is conceivable for example that neural networks will learn not only to compute decision-making functions but also to output why they have come to a specific decision.

So-called "recurrent neural networks" in particular, where information is not only propagated in one direction, have recently led to significant successes. We could think of these kinds of systems in simplified terms as neural networks with memory. These are in fact mathematically universal, which in theory means that there is no task that this kind of network could not learn and solve.

However, since recurrent neural networks simultaneously represent complex, nonlinear, dynamic feedback systems, their behavior can only be described mathematically with difficulty. This leads to general challenges with respect to the training process that can at present be circumvented by means of sheer computing power, meaning that training recurrent networks in practice is indeed possible. Nevertheless, research is going on into this issue across the globe and the pace of progress in this field leads us to expect that here, too, further breakthroughs will soon be achieved. We can therefore reckon on neural networks being increasingly involved in our professional and everyday lives in the near future; soon there will be few limits to the practical application of learning systems.

14.3 Cognitive robotics in production and services

Prof. Dr. Thomas Bauernhansl · Dipl.-Ing. Martin Hägele M.S. ·
Dr.-Ing. Werner Kraus · Dipl.-Ing. Alexander Kuss
Fraunhofer Institute for Manufacturing Engineering and Automation IPA

One area where machine learning (ML) techniques can be used is robotics. When we are talking about these machines, there is one traditional image that generally still comes to mind: a variety of robotic arms carrying out precise and strictly defined movements in large production facilities, for example for handling or welding purposes. This is where robots demonstrate their strengths, including repeat accuracy, precision, and the ability to handle heavy loads. Once programmed, they may continue carrying out this one activity for many years, paying off the significant initial investment in time and money by the long period of operation. Nevertheless, changes to the production process or parts entail programming outlay since the robot generally does not “know” how to interact with procedural changes. Its actions are thus limited in their autonomy.

Increasing autonomy is the goal currently being pursued via developments in so-called cognitive robotics: the aim is for robots, according to the term cognitive, to be able to demonstrate perception, recognition, and implement corresponding action by means of improved technologies. The basic principle here is that the system perceives something via its sensors, processes the data, and uses this to derive an appropriate response.

In the production environment, this allows users to respond flexibly as well as economically to the growing requirements of short product lifecycles and increasing numbers of variants. Capturing and analyzing large volumes of data in near real time with the aid of the machine learning processes explained in Chapter 14.2 enables systems to learn from the actions carried out and their success or failure.

For many of these systems, the basis for the degree of autonomy required is formed by cognitive functions since service robots must often navigate dynamic, potentially unknown environments. They need to be able to identify, pick, and pass objects, and even identify faces. The behavior for these actions cannot be preprogrammed, at least not in its entirety, and hence sensor information has to form the basis for adaptive decision-making. Only with cognitive functions robots are able to leave the production environment and become part of service provision in the commercial or home environments. A number of new areas of application have developed here in the last 20 years or so, including, in particular, agriculture, logistics, healthcare, and rehabilitation.

14.3.1 Intelligent image processing as a key technology for cost-efficient robotic applications

The primary domains for industrial robots (IRs) are in work piece handling (50% of all IRs) and welding (25% of all IRs) [source IFR]. In mass production, robot programs are mostly specified via manual teach-in processes and then carried out blindly, so to speak. For years now, image processing systems (computer vision) have been utilized e.g. for picking random objects such as in bin picking. Current solutions for object recognition utilize recognizable features and fixed programmed model-based object recognition methods. This means it poses no problem to the robot that the work pieces are randomly placed since the system can use its cognitive functions to identify where each work piece lies and how best to pick it.

Algorithm configuration as well as, for example, training for new work pieces is carried out manually by experts. Machine learning can now be utilized so that the bin picking algorithms – for object recognition and position estimation, for picking, or for manipulation for example – can be optimized autonomously on the basis of the analyzable information: the calculation times for picking shorten while the rate of successful picks rises. Process reliability thus increases with each picking attempt.

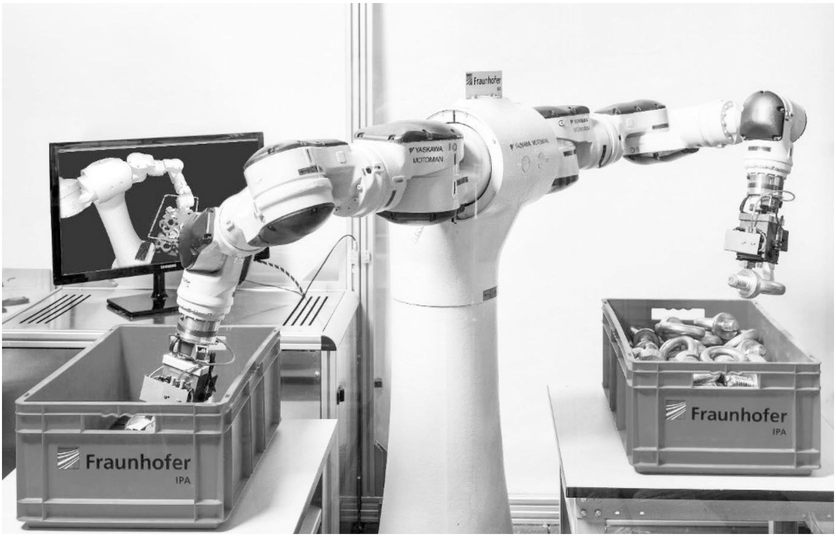


Fig. 14.3 Machine learning processes optimize robot-based systems for bin picking. (Fraunhofer IPA/photo: Rainer Bez)

As described in Chapter 14.2.1, training the machine learning process initially requires a large volume of training data. One typical approach is to generate the training data experimentally – i.e. in the case of bin picking, by carrying out and analyzing several hundred thousand picking attempts – before the neural networks can be utilized for stable operation. Since this time-consuming generation of training data is not practicable for industrial operation, a virtual learning environment is currently coming into being at Fraunhofer IPA in the form of a simulation model. Using this, numerous picking processes can be virtually carried out with the work piece required even before commissioning and without interrupting production. A so-called neural network – that is, a combination of connected processing units at various levels of abstraction (cf. Ch. 14.2.1) – learns from a large number of simulated picks and continuously improves its knowledge of the process. The pre-trained networks are then uploaded on the actual robots.

Welding robots will also in future be able to benefit from 3D sensors. To date, only about 20% of robotic systems utilize cameras to recognize work pieces for processing or picking and to plan actions accordingly, for example. It is possible to distinguish between two essential approaches to programming mechanically guided welding processes using robots:

- Teach-in programming, i.e. the robot program is generated by the operator with the actual work piece in the robot cell
- Offline programming based on the CAD model of the assembly with subsequent robotic search runs or partial re-teaching

In the case of offline programming, the robot is programmed in a virtual simulation environment without taking the actual robot cell out of action during the programming process. The key problem with offline programming is that differences arise between the virtual and actual production cells, e.g. due to work piece tolerances or deviations in component/peripheral component positions. This can then lead to insufficient or at least suboptimal weld quality. In practice then, the robot programs produced offline often need to be further adapted via teach-in programming. This can result in significant manual adjustments having to be carried out, especially in the case of changing production scenarios and varying component tolerances. By using optical 3D sensors and corresponding Fraunhofer IPA software, the robot is given the ability to “see” in a similar way to humans. In this way, the robot recognizes variations and is able to take account of these even during program planning. The robot is thus able to optimally adapt its behavior to changeable manufacturing conditions, as an error-tolerant production system.

Cognitive functions can also be utilized to simplify the programming process. Current programming systems require specialist operating skills and significant work to produce new robot programs. Particularly in mid-size companies, specially trained staff is often not available for this. The changing models and small batch sizes that are so common for mid-size firms additionally increase the amount of programming work and frequently impede the cost-effective utilization of robots. Here, software developed by Fraunhofer IPA makes it possible, for example, to automatically detect potential weld seams on work pieces using rules-based pattern recognition processes. The operator selects the seams to be welded and specifies the welding sequence. The time-consuming process of specifying coordinates to define weld seams is thus avoided and the programming process simplified.

Cognitive techniques can also be utilized in the field of collision-free pathway planning. The robot uses 3D sensors to capture its working environment and identifies objects posing a collision risk. Software developed by Fraunhofer IPA enables the automatic calculation of a collision-free pathway for the robot, using sample-based pathway planning processes, thus significantly reducing manual programming effort.

All of the data from the robotic system – for example, the position of the robot, sensor data, or operator inputs – is combined in a near real-time digital model. This forms the basis for the utilization of machine learning processes and additionally

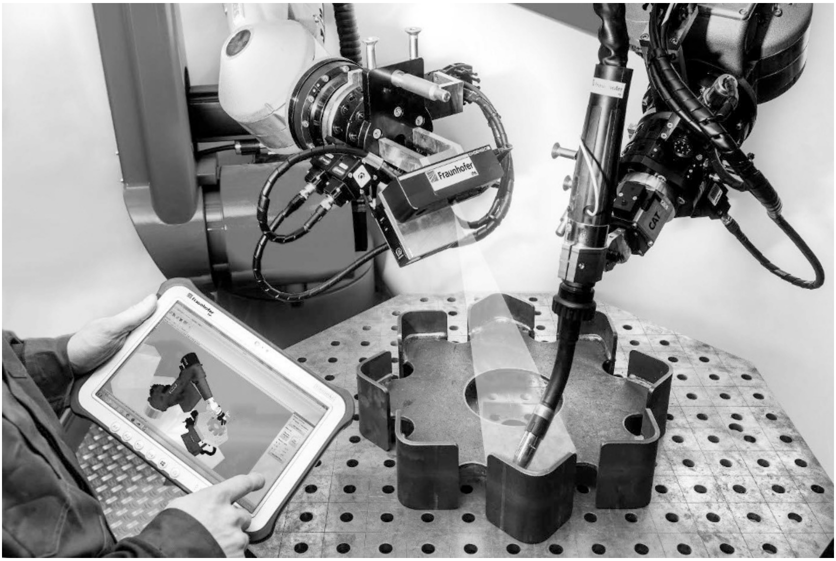


Fig. 14.4 With the welding robot software, welding even very small batches become economically viable using a robotic system. (Fraunhofer IPA/photo: Rainer Bez)

facilitates connection to a cloud infrastructure in order to implement analysis and learning processes across plants. In tests of the welding robot cell in a mid-size enterprise, the Fraunhofer IPA software reduced the programming time from 200 min. to 10 min. compared with manual teach-in processes, with the work required for training also being significantly reduced compared with existing programming processes.

14.3.2 A multifaceted gentleman: The Care-O-Bot® 4 service robot

For more than 20 years now, Fraunhofer IPA has been working on the vision of a service robot that can be used in numerous environments including hospitals and care facilities, warehouses, and even hotels. In 2015, researchers from Stuttgart presented the fourth generation model, developed in partnership with Schunk und Phoenix Design. Care-O-Bot® 4 offers diverse options for interaction, is highly agile, and allows for various system enhancements with its modular approach. This might start with the mobile platform for transportation, and then the torso-like structure can be used with one or two arms, or none at all. Care-O-Bot® 4 is equipped

Fig. 14.5 Sales clerk Paul accompanies customers to the product they are looking for in a Saturn store. (Image rights: Martin Hangen, right of reproduction: Media-Saturn-Holding)

with several sensors with which it is able to recognize its environment as well as objects and people and orient itself in space and even navigate freely. Researchers also placed an emphasis on a clear and appealing design.

The robot thus uses numerous cognitive functions with which it is able to perceive its environment and operate autonomously. This is illustrated by an active project with Media-Saturn-Holding: Care-O-Bot® 4 is being utilized in Saturn stores as a sales assistant, a "digital staff member" called Paul. It greets customers upon arrival, asks them which product they are looking for and accompanies them to the correct shelf. It is able to recognize objects and its environment, orienting itself and moving freely within it. Thanks to speech recognition software, it is able to conduct conversations with customers. To do this, Paul is equipped with domain-specific knowledge, that is, it is able to understand typical terms and topics within the specialist electronics trade (e.g. products, services, online order collections, etc.) and supply relevant information. In addition, it is able to appraise its dialog with customers and "understand" satisfied or critical feedback. And since it recognizes faces, it can tailor its communication to the customer's age and mood, for example. The goal of using the robot is to offer customers an innovative retail purchasing experience in store. It is additionally able to relieve staff by serving customers as their first port of call and assisting with finding products. For detailed queries, it then calls staff from customer services.

In future, it will be essential for cognitive robotics that knowledge acquired at one time is also made available centrally and can thus be used by several systems. The Paul in Ingolstadt, for example, will in future be able to share its acquired knowledge with the Paul in Hamburg via a private cloud. In as much as the complexity of what a robot needs to know is constantly increasing. In the same way, individual systems are already able to utilize knowledge available from online sources. Paul the robot is thus linked to Saturn's online store so it can utilize the product information already provided.

14.4 Off road and under water: Autonomous systems for especially demanding environments

Prof. Dr. Jürgen Beyerer
Fraunhofer Institute of Optronics, System Technologies,
and Image Exploitation IOSB

As explained above, the cognitive functionality requirements for autonomous robotic systems are significantly higher than those for conventional industrial robots. In order to be able to operate and fulfill tasks autonomously in an unknown and dynamic environment, the environment must be explored and suitably modeled, as outlined in Ch. 14.3. To do this, an “algorithm toolbox for autonomous robotic systems” was developed in modular form: it contains components for all processing stages ranging from environmental perception through task and motion planning to the final execution of this plan.

14.4.1 Autonomous mobile robots in unstructured terrain

Autonomous navigation requires several components. First, the platform must be equipped with sensors for localization and environmental perception. Using multi-sensor fusion, measurements from various sensors are combined to achieve a greater level of accuracy. Thus, for example, laser scanners and cameras are used to

Fig. 14.6 The IOSB.amp Q2 autonomous mobile robot is able to support rescue workers in quickly gaining an overview of the situation. (Fraunhofer IOSB)

produce a map of the environment so the robot can explore an unknown environment by itself. In order to constantly improve the localization and map via fusion with current sensor data, probabilistic methods for simultaneous localization and mapping are employed.

To allow collision-free motion planning, the sensor data is continuously used to generate a current obstacle map. It includes details on navigability and soil condition. The planning integrates both the map as well as the kinematic and dynamic properties of the platform in order to generate an optimal trajectory with guaranteed navigability. During planning, a variety of additional optimization criteria can be incorporated such as driving with maximum efficiency or preferably driving on even paths based on the data of the soil condition.

Due to the modular concept of the Algorithm Toolbox, a variety of different robot platforms can be equipped with the relevant autonomous abilities for their respective uses without requiring major adaptation efforts. The Algorithm Toolbox is thus utilized both for all-terrain platforms like the IOSB.amp Q2 (see Fig. 14.6) as well as for larger trucks.

14.4.2 Autonomous construction machines

The methods of cognitive robotics can also be used for mobile work machines. Examples include diggers and other construction machines, forklift trucks, and agricultural and forestry machines. These are usually operated manually and generally need to be converted for autonomous operation first. This includes facilities for electronically controlling driving and operational functions but also sensors for capturing the current state of the machine e.g. of the joint angle of the digger arm. It is also important to produce a computerized model of the work machine so that the joint angle measurement can be used to calculate where exactly the grab/excavation bucket or any other part of the machine is situated.

Once these conditions are in place, a mobile work machine can be equipped with autonomous capabilities from the Algorithm Toolbox like a robot. In comparison to mobile robots, additional methods for carrying out dedicated tasks with the manipulator are required here. A manipulator is any moving part that is capable of fulfilling handling tasks, such as the digger arm for example.

When using 3D environmental mapping, the fact that the manipulator may be in the sensor's field of view, requiring algorithms for distinguishing the manipulator and obstacles in the sensor data, must be considered. Based on the 3D environmental mapping and the geometric model of the machine, collision-free manipulator movements can be planned for the handling of tasks. The control of hydraulic driv-

Fig. 14.7 Fraunhofer IOSB autonomous digger (Fraunhofer IOSB)

en work machines is significantly more demanding than that of electrically operated industrial robots, because the system responds with a notable delay and load dependency.

At Fraunhofer IOSB, an autonomous mini digger serves as a demonstrator for mobile work machines (see Fig. 14.7). It allows, for example, for autonomous earth excavation within a predefined area selected by the user via a graphical interface. An additional potential usage scenario for autonomous mobile work machines is the recovery and removal of hazardous substances.

14.4.8 Autonomous underwater robots

Examples of underwater applications for autonomous robots include inspections and the exploration of the seabed. The demanding environment poses particular challenges to the cognitive functions of the robotic system. Integrating sensors for precise environmental imaging requires research into new, innovative carrier platforms. For this reason, between 2013 and 2016 a prototype for a new AUV (autonomous underwater vehicle) known as DEDAVE was designed, built, and tested at Fraunhofer IOSB, providing a range of benefits compared with similar underwater vehicles on the market (see Fig. 14.8). DEDAVE is light, exceptionally compact, and has a large maximum payload range. The vehicle can thus carry the complete sensor setup required for exploration tasks, something that is otherwise only possible with AUVs twice as big and heavy.

Fig. 14.8 DEDAVE autonomous underwater vehicle (Fraunhofer IOSB)

The vehicle carries the usual sensor systems for both mapping and sounding the seafloor simultaneously so that the various sensor data can be recorded without the need for surfacing, refitting, and submerging again. In this way, fewer missions are required and ship costs are reduced. Pressure-independent construction of suitable components reduces the vehicle weight and saves space that would otherwise be required for large pressure hulls. The new design of the rectangular payload section permits the application of current sonars and optical systems (required by potential customers). Water samplers and new systems for future applications can also be easily integrated here.

Extensive mission planning support is provided to the user by means of semiautomatic and automatic functions as well as libraries of driving maneuvers for the DEDAVE vehicle. DEDAVE's modular guidance software allows the user to change the AUV's behavior depending on sensor events. Since the patented system allows energy and data storage to be exchanged without the use of tools or lifting gear, very short turn-around times of approx. one hour are achievable between missions.

14.4.4 Summary

The modular concept of the components developed at Fraunhofer IOSB enables a broad range of robotic platforms for autonomous operation in a highly flexible way. Fraunhofer IOSB possesses a range of powerful technology demonstrators for testing and evaluating the modules developed.

14.5 Machine learning for virtual product development

Prof. Dr. Jochen Garcke · Dr. Jan Hamaekers · Dr. Rodrigo Iza-Teran
Fraunhofer Institute for Algorithms and Scientific Computing SCAI

Machine learning is being utilized increasingly often in materials research and product development to support the design engineer in the research and development process. Here, numerical simulations are currently used in many cases so that expensive and time-consuming actual experiments can be avoided, i.e. the fundamental technical and physical processes are calculated in advance on computer systems using mathematical-numerical models. Numerical simulations are used in the automotive industry to investigate the influence of different material properties, component shapes, or connecting components in various design configurations. In materials development and chemistry, multiscale modeling and process simulation are used to predict the properties of new materials even before they have been actually synthesized in the laboratory. This approach allows materials and manufacturing processes to be suitably optimized for specific requirements.

Efficient and data-driven work with large numbers of numerical simulations has so far only been possible to a limited degree. That is to say, nowadays comparisons of the different results examine only a small number of key quantities but not the detailed outcomes of the highly complicated actual simulations, for example the different deformations. Against this backdrop new methods of machine learning are being developed and applied at Fraunhofer SCAI for analyzing, utilizing, and processing results data from numerical simulations [13] [14].

14.5.1 Researching crash behavior in the automotive industry

Virtual product development in the automotive industry utilizes numerical simulations to analyze, for example, the crash behavior of different design configurations. Here, variations are made to material properties or component shapes, among other factors. Efficient software solutions are in place to assess several simulation results as long as only simple parameters such as bends, intrusion of the firewall at selected points, acceleration, etc. are being investigated. For detailed analysis of individual crash simulations, specialized 3D visualization software is utilized.

In order to analyze this large amount of complex data, we use machine learning (ML) approaches for so-called nonlinear dimensionality reduction. With this a low-dimensional representation is calculated from the available data. By visually arranging the data with regard to this small number of key quantities calculated using ML approaches, a simple and interactive overview of the data (in this case, the simulation results) is facilitated. In particular, we have developed a method which calculates a small number of elemental and independent components from the data and thus enables the representation of a numerical simulation as their combination. This data-based decomposition can be understood as a kind of elemental analysis of component geometries and facilitates a very compact and efficient depiction. Elemental modes obtained when examining crash simulations may include for example the rotation of a component or its global or local deformation in one of the component's areas, which especially also enables a physical interpretation of the analysis results. In this way, a study can be carried out efficiently since all of the simulations can be described with the aid of these elemental components and thereby compared.

This reduction in data dimensionality facilitates the intuitive interactive visualization of a large number of simulations. An interpretable arrangement in three coordinates with respect to selected elemental analyses demonstrates the differences between the simulations, for example the various geometric deformations during the crash. As an example, we will take a look at a digital finite element model of a pickup truck, which we studied in the BMBF big data project, VAVID [14]. A frontal crash is simulated, with the component plate thicknesses being varied. The deformations of the main chassis beam are the subject of analysis. Due to the new representation by the calculated elemental decomposition, it is possible to represent the various deformations as the sum of the elemental components in a manner that is compact and interpretable.

For each simulation, we examine around a hundred intermediate steps that are then visualized simultaneously in a graphic, something that was impossible with previous analysis methods. By means of these components, our ML methods allow the progression of the crash behavior over time to be neatly illustrated (see Fig. 14.9). Each point here represents a simulation for a specific time step. It can be clearly seen how all of the simulations begin with the same geometry and over time produce two variations of the crash behavior, illustrated in each case by means of

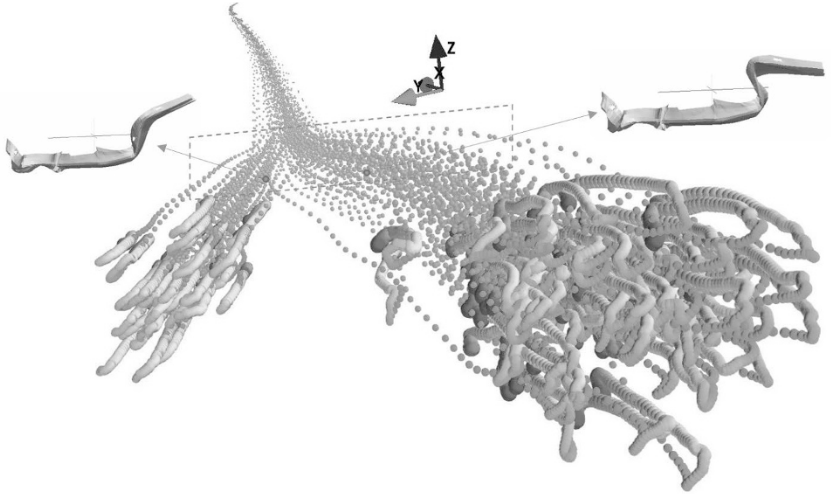


Fig. 14.9 Comparative analysis of around 100 time-dependent simulations (Fraunhofer SCAI)

typical deformations of the main chassis beam under observation. In addition, the point in time at which this division occurs can be approximately identified. On the basis of this kind of analysis of the simulation results, the development engineer is better able to decide how design parameters should be chosen.

In addition to this, new digital measuring techniques have also been developed and made available in recent years that enable high-definition, time-dependent 3D data to be obtained from actual crash tests. The techniques that we have newly developed facilitate comparisons to be carried out for the first time between simulations and the precise measurement data from an actual experiment [14]. In this way, the correct numerical simulation for an actual crash test can be identified, something that was previously not feasible at this level of quality. This also makes it possible to obtain an overview of all of the simulations and identify whether an actual experiment is proceeding along the left or right deformation path as per Fig. 14.9.

These recently developed ML methods are able to significantly simplify and accelerate the virtual product development R&D process since the development engineer requires less time for data preparation and analysis and can concentrate on the actual core technical engineering tasks.

14.5.2 Designing materials and chemicals

The full range of materials, chemicals, and active agents possible is absolutely vast. The number of active agent molecules alone, for example, is estimated at 10^{60}, with the number of molecules with 30 atoms or less lying between 10^{20}–10^{24}. By contrast, fewer than 100 million known stable compounds are currently accessible in public databases. The difference between known and potential compounds suggests that a very large amount of new materials, chemicals, and active agents likely remain to be discovered and developed. On the other hand, the huge extent of this scope represents an enormous challenge for the design of new materials and chemicals since exploring it is usually very costly. Fraunhofer SCAI is particularly researching special ML methods here to substantially accelerate these kinds of design and optimization processes for materials and chemicals.

For example, numerous materials and molecule databases are currently being built worldwide, with the results of quantum chemical simulations in particular. These can be used to develop efficient ML-based models for predicting properties. A model developed by Fraunhofer SCAI can thus be trained on a multitude of smaller molecules, but utilized especially for predicting molecules of arbitrarily greater sizes. Here, the techniques developed allow chemical accuracy to be achieved for many properties [13]. In this way, the costs of an elaborate quantum chemical calculation can be substantially reduced by several orders of magnitude (typically from a few hours to a few milliseconds). The descriptors and distances specially designed by Fraunhofer SCAI to be suitable for molecules and materials can also be used for ML and analysis methods to identify interesting and promising areas within the range of all compounds for more precise exploration.

Overall, by means of ML techniques and numerical simulations the development and design of new materials and chemicals can be made significantly simpler, faster, and more cost effective.

Sources and literature

[1] Bengio Y (2009) Learning Deep Architectures for AI, Foundations and Trends in Machine Learning 2 (1) 1-127

[2] Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet Classification with Deep Convolutional Neural Networks, in *Proc. NIPS*

[3] Hinton G, Deng L, Yu D, Dahl G, Mohamed A and Jaitly N (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Processing Magazine* 29 (6) 82-97

[4] Mikolov T, Sutskever I, Chen K, Corrado G and Dean J (2013) Distributed Representations of Words and Phrases and their Compositionality in *Proc. NIPS*

[5] Levine S, Wagener N, Abbeel P (2015) Learning Contact-Rich Manipulation Skills with Guided Policy Search, Proc. IEEE ICRA

[6] Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, Thrun S (2017) Dermatologist-level Classification of Skin Cancer with Deep Neural Networks, *Nature* 542 (7639) 115-118

[7] Silver D, Huang A, Maddison C, Guez A, Sifre L, van den Driesche G, Schrittwieser J (2016) „Mastering the Game of Go with Deep Neural Networks and Tree Search" *Nature* 529 (7587) 484-489

[8] Moracik M, Schmidt M, Burch N, Lisy V, Morrill D, Bard N, Davis T, Waugh K, Johanson M and Bowling M (2017) DeepStack: Expert-level Artificial Intelligence in Heads-up No-limit Poker, *Science* 356 (6337) 508-513

[9] Hornik K, Stinchcombe M, White H (1989) Multilayer Feedforward Networks Are Universal Approximators, *Neural Networks* 2 (5) 359-366

[10] Cybenko G (1989) Approximation by Superpositions of a Sigmoidal Function, *Mathematics of Control, Signals and Systems* 2 (4) 303-314

[11] Rumelhart D, Hinton G and Williams R (1986) Learning Representations by Back- propagating Errors, *Nature* 323 (9) 533-536

[12] V. Vapnik and A. Chervonenkis (1971) On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, *Theory of Probability and its Applications* 16 (2) 264-280

[13] Barker J, Bulin J, Hamaekers J, Mathias S, LC-GAP (2017) Localized Coulomb Descriptors for the Gaussian Approximation Potential, in Scientific Computing and Algorithms in Industrial Simulations, Griebel, Michael, Schüller, Anton, Schweitzer, Marc Alexan- der (eds.), Springer

[14] Garcke J, Iza-Teran R, Prabakaran N (2016) Datenanalysemethoden zur Auswertung von Simulationsergebnissen im Crash und deren Abgleich mit dem Experiment, *VDI-Tagung SIMVEC 2016*.

15

Fraunhofer Big Data and Artificial Intelligence Alliance

Mining valuable data

Prof. Dr. Stefan Wrobel · Dr. Dirk Hecker
Fraunhofer Institute for Intelligent Analysis and
Information Systems IAIS

Summary

Big data is a management issue across sectors and promises to deliver a competitive advantage via structured knowledge, increased efficiency and value creation. Within companies, there is significant demand for big data skills, individual business models, and technological solutions.

Fraunhofer assists companies to identify and mine their valuable data. Experts from Fraunhofer's Big Data and Artificial Intelligence Alliance demonstrate how companies can benefit from an intelligent enrichment and analysis of their data.

15.1 Introduction: One alliance for many sectors

The data revolution will bring lasting and massive changes to many sectors. As far back as 2013, this development was identified by the German Federal Ministry for Economic Affairs when they commissioned Fraunhofer to conduct an initial analysis of the use and potential of big data in German companies [14]. The aim was to highlight potential courses of action for economy, politics, and research. Extensive web research, a survey and several sector workshops soon revealed that companies could be more efficiently managed through real-time analysis of their process and business data, that analyzing customer data could facilitate increasingly personalized services, and that connected products could be equipped with a greater level of

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_15

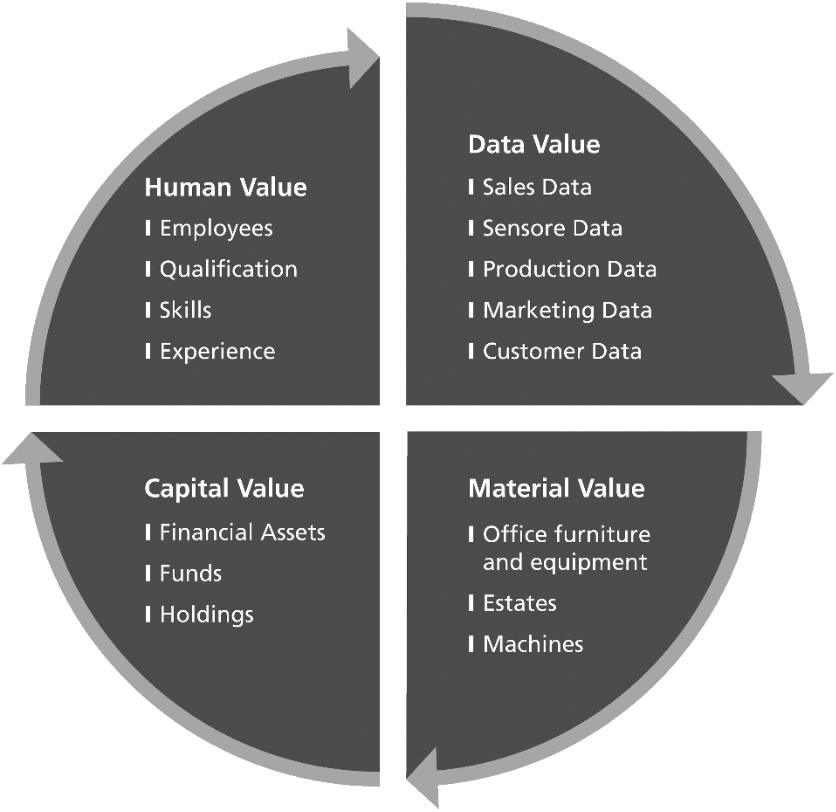


Fig. 15.1 Data forms a new pillar of company value. (Fraunhofer IAS)

intelligence. A company's data becomes a hard-to-replicate competitive advantage if it can profitably be analyzed and harnessed for the company's activities, services, and products. However, according to the recommendations by Fraunhofer experts during the market analysis, companies must invest in the exploitation and quality of their data. They must develop ideas for how they can best use the data for their business [5]. And they must develop corresponding in-house expertise. This is illustrated in Fig. 15.1: big data forms a new category of company value and is thus a key management issue.

The task for the Fraunhofer Big Data and Artificial Intelligence Alliance, formed shortly after the study's publication, was thus clear: supporting companies on their journey towards becoming data-driven businesses. The Alliance currently offers direct access to the diverse skills provided by Fraunhofer experts from more than 30 institutes. It thus combines unique sector-specific know-how across Germany with deep knowledge of recent research methods in intelligent data analysis. The benefit for companies? Best practices and use cases from other sectors can easily be adapted and utilized for creative solutions.

The box that follows provides an overview of the Alliance's core areas of business.

Fraunhofer Big Data and Artificial Intelligence Alliance core business areas

Production and industry: This is concerned with making better use of the growing volumes of data in production. By utilizing machine learning methods, processes can be optimized, errors recognized sooner by means of anomaly detection, and human-machine interaction made safer by means of improved robotics technology.

Logistics and mobility: The core issue of this business area is optimizing the whole of logistics across all modes of transport. This helps to reduce empty runs, waiting times, and traffic jams. Autonomous vehicles improve traffic flow and provide occupants with opportunities to make good use of travel time.

Life sciences & health care: Connected devices monitor patients during their everyday lives, intelligent systems evaluate medical data autonomously, and telepresence robots facilitate distance diagnoses. The use of modern data analysis in life sciences and health care offers numerous options and opportunities for the future.

Energy & environment: Data can be used to predict noise in cities and identify sources of noise, to analyze flora and fauna, and to optimize the energy management of large buildings.

Security: Using pattern recognition technology and independent learning capabilities, security systems can massively be improved and cyber defense systems made even more precise.

Business and finance: Adaptive billboards, question answering systems, and deep learning methods in text recognition – the potential applications for data analysis in business are numerous.

Sources and literature

[1] **Auffray** C, Balling R, Barroso I et al (2016) Making sense of big data in health research: Towards an EU action plan. Genome Medicine. URL: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0323-y

[2] **Cacilo** A, Schmidt S, Wittlinger P et al (2015) Hochautomatisiertes Fahren auf Autobahnen – Industriepolitische Schlussfolgerungen. Fraunhofer IAO. URL: http://www.bmwi.de/Redaktion/DE/Downloads/H/hochautomatisiertes-fahren-auf-autobahnen.pdf?__blob=publicationFile&v=1

[3] **Chemnitz** M, Schreck G, Krüger J (2015) Wandlung der Steuerungstechnik durch Industrie 4.0. Einfluss von Cloud Computing und Industrie 4.0 Mechanismen auf die Steuerungstechnik für Maschinen und Anlagen. Industrie 4.0 Management 31 (6): 16-19

[4] **Fedkenhauer** T, Fritzsche-Sterr Y, Nagel L et al (2017) Datenaustausch als wesentlicher Bestandteil der Digitalisierung. PricewaterhouseCoopers. URL: http://www.industrialdataspace.org/ressource-hub/publikationen/

[5] **Fraunhofer IGB** (2016) Mass Personalization. Mit personalisierten Produkten zum Business to User (B2U). Fraunhofer IGB. URL: https://www.igb.fraunhofer.de/de/presse-medien/presseinformationen/2016/personalisierung-als-wachstumstreiber-nutzen.html

[6] **Fraunhofer IVI**, Quantic Digital, Verbundnetz Gas AG (VNG AG) (2016) Wegweiser zum Energieversorger 4.0: Studie zur Digitalisierung der Energieversorger – In 5 Schritten zum digitalen Energiemanager. Quantic Digital. URL: www.quantic-digital.de/studie-digitalisierung-energieversorger/

[7] **Heyen** N (2016) Digitale Selbstvermessung und Quantified Self. Potenziale, Risiken und Handlungsoptionen. Fraunhofer ISI. URL: http://www.isi.fraunhofer.de/isi-wAssets/docs/t/de/publikationen/Policy-Paper-Quantified-Self_Fraunhofer-ISI.pdf

[8] **Klink** P, Mertens C, Kompalka K (2016) Auswirkung von Industrie 4.0 auf die Anforderungen an ERP-Systeme. Fraunhofer IML. URL: https://www.digital-in-nrw.de/files/standard/publisher/downloads/aktuelles/Digital%20in%20NRW_ERP-Marktstudie.pdf

[9] **Radić** D, Radić M, Metzger N et al (2016) Big Data im Krankenversicherungsmarkt. Relevanz, Anwendungen, Chancen und Hindernisse. Fraunhofer IMW. URL: https://www.imw.fraunhofer.de/content/dam/moez/de/documents/Studien/Studie_Big%20Data%20im%20Krankenversicherungsmarkt.pdf

[10] **Schmidt** A, Männel T (2017) Potenzialanalyse zur Mobilfunkdatennutzung in der Verkehrsplanung. Fraunhofer IAO. URL: http://www.iao.fraunhofer.de/lang-de/images/iao-news/telefonica-studie.pdf

[11] **Steinebach** M, Winter C, Halvani O et al (2015) Chancen durch Big Data und die Frage des Privatsphärenschutzes. Fraunhofer SIT. URL: https://www.sit.fraunhofer.de/fileadmin/dokumente/studien_und_technical_reports/Big-Data-Studie2015_FraunhoferSIT.pdf

[12] **Urbach** N, Oesterle S, Haaren E et al (2016) Smart Data Transformation. Surfing the Big Wave. Infosys Consulting & Fraunhofer FIT. URL: http://www.fit.fraunhofer.de/content/dam/fit/de/documents/SmartDataStudy_InfosysConsulting_FraunhoferFIT.pdf

15.2 Offerings for every stage of development

Companies that turn to the Fraunhofer Big Data and Artificial Intelligence Alliance are at different stages in their journeys to digitization and benefit from a modular offering divided into four levels, as shown in Fig. 15.2.

In the beginning, the focus is on getting to know the potential in your own sector, generating enthusiasm among staff, and coming up with initial ideas for your own company. This is facilitated by excite seminars and innovation workshops. The guiding principle here is to learn from the best.

At the next level, the aim is for the most promising ideas to gain momentum quickly and efficiently. A data asset scan is used to identify all of the relevant internal data, and a search is conducted for complementary publicly available data. The most important tools and algorithms are integrated within a Fraunhofer starter toolkit so that even large datasets can be analyzed quickly. Fraunhofer institutes provide support where needed in developing concepts, demonstrators, and scalable prototypes.

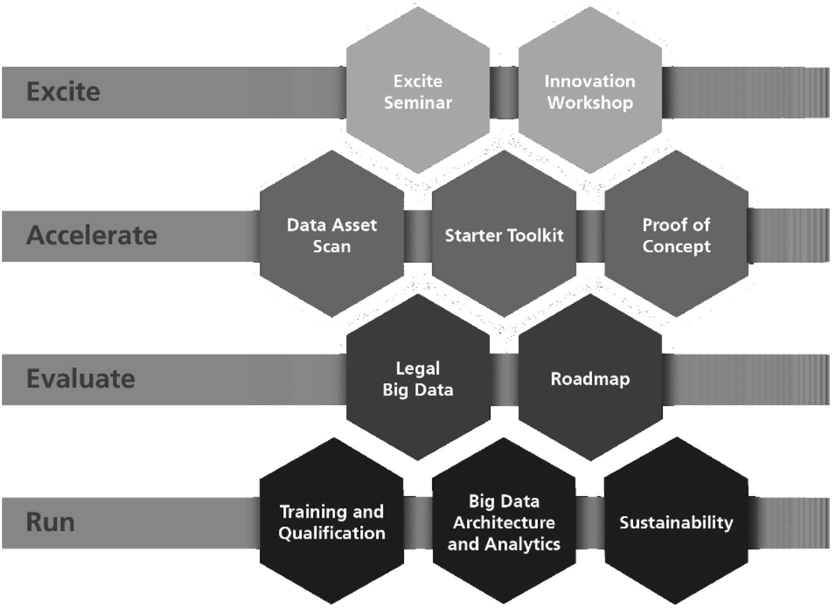


Fig. 15.2 Four levels of becoming a data-driven company (Fraunhofer IAIS)

After their initial practical experiences, it is helpful for companies to reflect on the path they are pursuing. How should the use cases be assessed between the priorities of technical feasibility, the operational context, and commercial attractiveness? What might a tailored big data strategy and roadmap look like? Can personal data be utilized legally by applying principles of "privacy by design"?

To get companies started, Fraunhofer institutes provide advice on choosing a suitable big data architecture and integrating analyses into operative processes. Fraunhofer's technology monitoring ensures that none of the important technological trends of the sector are missed. Guidelines for best practices and team mentoring guarantee the quality and efficiency of analysis projects.

The following sections take a look at data and people as company values, from the big data perspective: opportunities for data monetization, machine learning as the core technology of data analysis, and training data scientists as experts for analysis.

15.3 Monetizing data

Prof. Dr. Henner Gimpel
Fraunhofer Institute for Applied Information Technology FIT

Data is a production and competitive factor and thus an asset. For this reason, it is worthwhile for companies to collect data or buy it in, to curate it, to protect it, and to intelligently combine and analyze it. A modern car, for example, is not merely a vehicle for transport, it can also be used as a weather station. From integrated thermometers and rain sensors or by logging the activation of windshield wipers, from light sensors, hygrometers and GPS, important data can be collected and transmitted via the mobile telephony network. By aggregating the data from a large number of vehicles and comparing this with satellite images, a very high-resolution local image of the current weather can be produced without substantial investment in infrastructure. Initial pilot trials have shown that there is significant willingness to pay for these kinds of local weather reports: traditional weather services may use them, for example, to optimize weather forecasting for agriculture or to estimate the power injection from photovoltaic and wind power facilities. A less obvious and new group of customers are facility managers, for example, who are able to use the data to optimize the air conditioning for large buildings and thus improve the indoor climate and energy efficiency.

Other forms of monetization are more subtle, as can be seen from a number of technology start-ups in the financial services sector – so-called fintechs. A research project conducted by Gimpel et al. [4] on the business models and service offerings of 120 fintech start-ups showed that for fintechs the rich data generated at the customer interface is an important production and competitive factor. Alongside data from individual customers, this increasingly includes comparisons with other customers (peers) and combinations with publicly available data.

Finally, monetization addresses the question of revenue model. There are various "currencies" that the customer may pay with. In only around a third of the cases they pay with money. In the largest number of service offerings observed, the users "pay" with loyalty, attention, or data. Loyalty to a brand or firm may lead to indirect monetization in other areas of operations. Both, selling data and advertising lead to financial payments from business partners. Where monetization via business partners is concerned, the service user is no longer the customer but the product itself.

The key dimensions for designing data monetization are thus immediacy (direct vs. indirect sales), the user's currency (money, data, attention, loyalty), and the role of the user (customer or product). Additional dimensions are the billing increment (subscription or transaction-based), price differential (none, segment-oriented, versioned, personalized), and price calculation (based on marginal costs that tend towards zero, or value for the user and customer).

Not all of the options that could ultimately be imagined for the use and monetization of data are legal and legitimate. There are substantial legal restrictions on the use of data (e.g. in Germany the Teleservices Act, Unfair Competition Act, EU General Data Protection Regulation), there are limits of acceptance on the part of users, customers, and partners, and there are ethical boundaries to be observed. Value-centered corporate leadership requires balancing these boundaries with the technological and economic possibilities and finding a viable middle way that creates value from data as a production and competitive factor and asset. In the future, users and regulators will also increasingly become aware of the value of data and its systematic analysis. More and more, users will demand appropriate recompense (free services, fees) for their data. Discussions are ongoing in specialist legal and political circles as to whether and how the value of data should be accounted for and assessed, and whether payments with data for allegedly free services should not be seen as payment as much as money. This would have clear consequences for the liability of providers of allegedly free services.

15.4 Mining valuable data with machine learning

Dr. Stefan Rüping

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

A few years ago, many companies were able to increase their efficiency simply by combining their data from previously separate “data silos” and analyzing it in a new way.

But the new possibilities of predictive and prescriptive analysis for early detection and targeted actions were soon recognized. Now, the trend is increasingly towards self-regulating – i.e. autonomous – systems [3]. Here, machine learning methods are being utilized that are particularly well applicable to large datasets. These methods identify patterns from historical data and learn complex functions that are also able to interpret new data and make proposals for action. These are mostly functions that could not be programmed explicitly due to the large number of different cases.

This becomes most evident in speech and image processing. Here, deep learning methods have recently achieved spectacular results by training networks with many layers of artificial neurons. They enable intelligent machines to speak with us in any language of choice and to perceive and interpret our shared environment [12]. Artificial intelligence creates a new communications interface with our home, car, and wearables, and is replacing touchscreens and keyboards. Major technology giants are spending up to $27 billion on internal research and development for intelligent robots and self-learning computers, and leading Internet giants are positioning themselves as companies for artificial intelligence [2]. Germany, having won a leading role for itself in the field of Industry 4.0, is now seeing the next wave of data rolling in with the Internet of Things [9]. IDC estimate that there will be 30 billion connected devices by 2020 and 80 billion by the year 2025 [7]. By using data generated during industrial manufacturing and over the entire product lifecycle, “industrial analytics” reveals potential for value creation along the entire production chain – from construction through logistics and production to reengineering. In the digital factory, where smart and connected objects constantly map the current state of all processes at all levels of production and logistics to a digital twin, demand-driven autonomous control of production processes, data-driven forecasting and decision-making become possible [5]. Work in the area of production is becoming more user-friendly and efficient because of intelligent assistance systems [1]. The use of machine learning techniques, especially in industry, requires high levels of domain-specific knowledge. This begins with the selection of the relevant data in the flood of sensors and notifications, its semantic enrichment, and the interpretation of the patterns and models learned. Here we are increasingly concerned with the combination of engineering and machine knowledge, with comprehensibility and liability, with the controllability of autonomous intelligent machines and collaborative robots, with data protection, with security and certifiability, with the fear of jobs losses and the need for new staff qualifications.

15.5 Data scientist – a new role in the age of data

Dr. Angelika Voß
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

The arrival of big data gave rise to the new occupational profile of data scientist [12]. Ideally, these data scientists are not only familiar with up-to-date software for

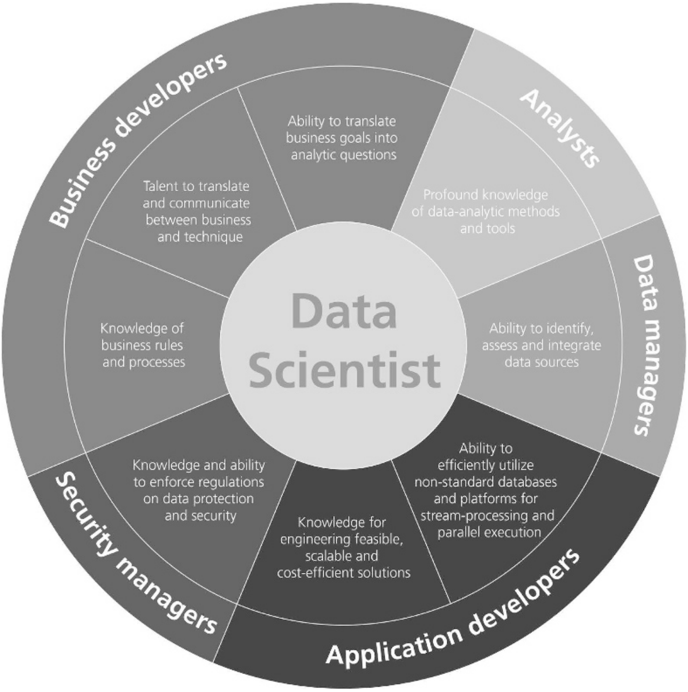


Fig. 15.3 Teams of data scientists bring together a number of talents. (Fraunhofer IAIS)

big data systems and predictive data analysis, but also understand the respective business sector and have the relevant domain knowledge.

As far back as 2011, McKinsey already warned that the need for specialists in big data analytics would become a bottleneck in companies in view of the rapid Increase in data volumes [11]. The growing need for processing images and videos, speech and text for digital assistants and intelligent devices is making the situation even more acute. Data scientists who master current tools such as Spark and Python are therefore well remunerated [8]. In practice, therefore, all-rounders are less common than teams with a mix of specialists.

Although many British and American universities offer masters courses for data scientists, and a number of German universities have now followed suit, this is not enough. Existing staff need to be trained for new methods, tools, languages, libraries and platforms, and sector specialists need to be able to collaborate in teams with data scientists. Here, too, we cannot wait until data science courses become established in the universities' curricula in the application sectors.

Fraunhofer Big Data and Artificial Intelligence Alliance has thus offered courses in big data and data analysis for professionals ever since it was founded. More specific training courses in the different tasks of data scientists followed: potential analysis for big data projects, security and data protection for big data, and – most recently – deep learning. Sector-specific and application-specific training courses followed, e.g. for smart energy systems, smart buildings, linked data in enterprises, data management for science, and social media analysis. In 2015, a three-level certification program for data scientists was started.

The Fraunhofer Big Data and Artificial Intelligence Alliance also advises companies that have their own professional development academies. Here, support ranges from the development and adaptation of modules to training in-house trainers in open Fraunhofer training courses. In 2016, for example, the Verband Deutscher Maschinen- und Anlagenbau (*The Mechanical Engineering Industry Association*, VDMA) produced plans for future machine learning in machinery and plant building [9] and is now offering online training materials developed for engineers, both with the support of Fraunhofer [15].

15.6 Conclusion

As far as value creation from data is concerned, US Internet giants followed by their Chinese counterparts are in the lead. They have the data, the platforms for big data processing and learning, they develop new methods, and by now they can draw on extensive trained models for speech and image comprehension. Industrial analytics and machine learning for intelligent machines and devices, however, offer a field where German industry and research can and must develop unique selling points with secure and transparent solutions. If production companies can use qualified staff and the right technologies to strategically mine their valuable data, then they will also be able to secure a lead in the face of international competition.

Sources and literature

[1] Acatech (2016) Innovationspotenziale der Mensch-Maschine-Interaktion (acatech IMPULS). Herbert Utz Verlag. München

[2] Bughin J, Hazan E, Ramaswamy S et al (2017) Artificial intelligence: the next digital frontier?. McKinsey Global Institute. URL: https://www.mckinsey.de/wachstumsmarkt-kuenstliche-intelligenz-weltweit-bereits-39-mrd-us-dollar-investiert (abgerufen: 07/2017)

[3] Böttcher B, Klemm D, Velten C (2017) Machine Learning im Unternehmenseinsatz. CRISP Research. URL: https://www.crisp-research.com/publication/machine-learning-im-unternehmenseinsatz-ku%CC%88nstliche-intelligenz-als-grundlage-digitaler-transformationsprozesse (abgerufen: 07/2017)

[4] Gimpel H, Rau D, Röglinger M (2016) Fintech-Geschäftsmodelle im Visier. Wirtschaftsinformatik & Management 8(3): 38-47

[5] Hecker, D, Koch D J, Werkmeister C (2016) Big-Data-Geschäftsmodelle – die drei Seiten der Medaille. Wirtschaftsinformatik und Management, Heft 6, 2016

[6] Hecker D, Döbel I, Rüping S et al (2017) Künstliche Intelligenz und die Potenziale des maschinellen Lernens für die Industrie. Wirtschaftsinformatik und Management, Heft 5, 2017

[7] Kanellos M (2016) 152 000 Smart Devices Every Minute In 2025: IDC Outlines The Future of Smart Things. Forbes. URL: https://www.forbes.com/sites/michaelkanellos/2016/03/03/152000-smart-devices-every-minute-in-2025-idc-outlines-the-future-of-smart-things/#2158bac74b63 (abgerufen: 07/2017)

[8] King J, Magoulas R (2017) 2017 European Data Science Salary Survey – Tools, Trends, What Pays (and What Doesn‘t) for Data Professionals in Europe. O’Reilly. URL: http://www.oreilly.com/data/free/2017-european-data-science-salary-survey.csp (abgerufen: 07/2017)

[9] Lueth K, Patsioura C, Williams Z et al (2016) Industrial Analytics 2016/2017. The current state of data analytics usage in industrial companies. IoT Analytics. URL: https://digital-analytics-association.de/dokumente/Industrial%20Analytics%20Report%202016%202017%20-%20vp-singlepage.pdf (abgerufen: 07/2017)

[10] Maiser E, Schirrmeister E, Moller B et al (2016) Machine Learning 2030. Zukunftsbilder für den Maschinen- und Anlagenbau. Band 1, Frankfurt am Main

[11] Manyika J, Chui M, Brown B et al (2011) Big data: The next frontier for innovation, competition, and productivity. McKinsey&Company. URL: http://www.mckinsey.com/

business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (abgerufen: 07/2017)
[12] Neef A (2016) Kognitive Maschinen. Wie Künstliche Intelligenz die Wertschöpfung transformiert. Z_punkt. URL: http://www.z-punkt.de/de/themen/artikel/wie-kuenstliche-intelligenz-die-wertschoepfung-treibt/503 (abgerufen: 07/2017)
[13] Patil D, Mason H (2015) Data Driven – Creating a Data Culture. O'Reilly. URL: http://www.oreilly.com/data/free/data-driven.csp (abgerufen: 07/2017)
[14] Schäfer A, Knapp M, May M et al (2013) Big Data – Perspektiven für Deutschland. Fraunhofer IAIS. URL: https://www.iais.fraunhofer.de/de/geschaeftsfelder/big-data-analytics/referenzprojekte/big-data-studie.html (abgerufen: 07/2017)
[15] University4Industry (2017) Machine Learning – 1.Schritte für den Ingenieur. URL: https://www.university4industry.com (abgerufen: 07/2017)

16

Safety and Security

Cybersecurity as the basis for successful digitization

Prof. Dr. Claudia Eckert
Fraunhofer Institute for Applied and Integrated Security AISEC
Prof. Dr. Michael Waidner
Fraunhofer Institute for Secure Information Technology SIT

Summary

Cybersecurity is the basis for successful digitization and for innovation in all sectors, e.g. in digital production (Industry 4.0), smart energy supply, logistics and mobility, healthcare, public administration, and cloud-based services, too. The role of cybersecurity [13][11] is to protect companies and their values and to prevent damage or at least limit the impact of any potential damage. Cybersecurity encompasses measures to protect IT-based systems (hardware and software) from manipulation and thus safeguards their integrity. Furthermore, it includes concepts and processes that guarantee the confidentiality of sensitive information and the protection of the private sphere as well as the availability of functions and services. Guaranteeing integrity, confidentiality, and availability are the familiar safety objectives already pursued by traditional IT security, but achieving them has become increasingly difficult and complex with digitization and networking and the accompanying connection between the digital and physical worlds.

The article that follows provides an insight into current trends and developments in the field of application-oriented cybersecurity research and makes use of selected example applications to outline challenges and potential solutions.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_16

16.1 Introduction: Cybersecurity – The number one issue for the digital economy

Cybersecurity is essential for sustainable value creation and for viable future business models. Current studies confirm this:

In mid-2017, the Bundesdruckerei published the findings of a representative survey [1] on IT security that highlights the importance of IT security for digitization: "Nearly three quarters of respondents view IT security as the foundation for successful digitization." In a survey of specialists conducted in early 2017 [2], the majority of respondents assessed the current significance of IT security for value creation in Germany as high, and nearly 90% were of the opinion that this significance would further increase over the next five years. The study commissioned by the German Federal Ministry for Economic Affairs and Energy stresses that IT security is not only an important market in itself, it is also the prerequisite for the development of additional viable future business models.

For the Bitkom industry association, IT security is also one of the two top issues in 2017 [3]: "IT security is becoming even more important since digitization is leading to more and more critical systems such as vehicles, medical technology, and machines becoming digitally connected," stated Bitkom General Manager Dr. Bernhard Rohleder. "At the same time, the attacks of criminal hackers are becoming increasingly refined. Regular security tools such as virus scanners and firewalls are often no longer enough for companies."

16.2 The (in-)security of current information technology

Digitization and networking lead to the development of complex cyber-physical systems where the boundaries between the digital and physical worlds disappear. Today, even the smallest sensors are able to capture a range of data and transfer this to cloud-based platforms extremely quickly, even across great distances. Data from very different sources is, for example, automatically analyzed using machine learning techniques and used to develop forecasts and recommendations for action. Increasingly, this data is being used to autonomously control critical processes, such as in automation engineering, in the operation of critical infrastructure, but also for autonomous vehicles. On top of that, the processed data often contains company-relevant know-how such as details of production processes that may only be passed on and used under controlled conditions. Digitization can thus only be secure if it is possible to guarantee the trustworthiness and reliability of all of the components involved in data processing and storage such as embedded sensors, cloud platforms or apps, as well as the machine learning processes utilized [14][15].

The problem, however, is that networking and digitization not only drastically raise the potential impact of successful cyberattacks, but that fundamental rethinking about our interaction with cyberthreats is also required, since the attack landscape, too, has changed dramatically in recent years. Cybercrime and cyberespionage have become professionalized [16]. Attacks are increasingly targeted at specific organizations and individuals and evade regular protective mechanisms such as firewalls, antivirus programs, and intrusion detection systems.

Today's attackers generally use human error, weak points in IT or in IT-based processes, and inadequate precautions in order to penetrate systems, manipulate them, or gain unauthorized access to sensitive information. One problem here is the ever-increasing number of hackers carrying out targeted attacks for their own benefit or to the disadvantage of a third party. A key aim of these kinds of attacks is to infiltrate systems with malicious software, so-called trojans, that gathers information such as passwords or login details unnoticed and makes them available to the attacker. The ease with which attackers with even limited technical knowledge can carry out these kinds of attacks represents an enormous threat for current connected

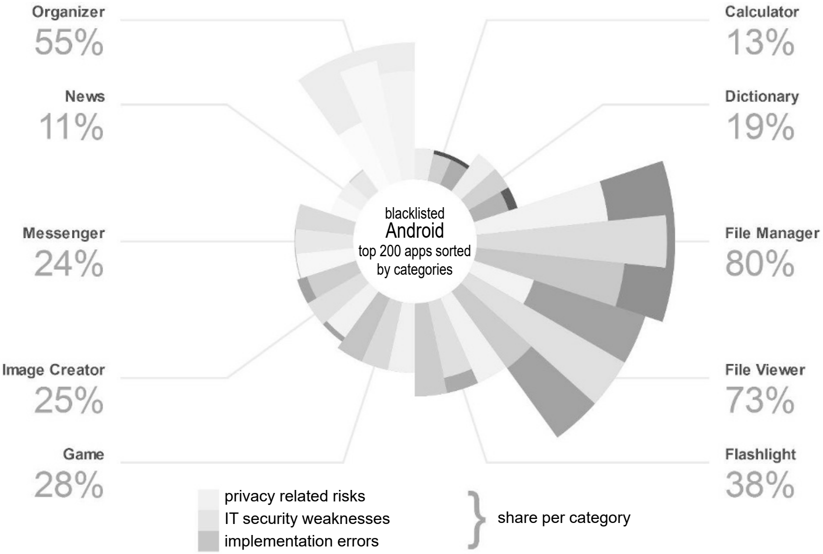


Fig. 16.1 Vulnerabilities of Android apps by application area according to the Appicaptor Security Index (in illustration of [6]) (Fraunhofer SIT)

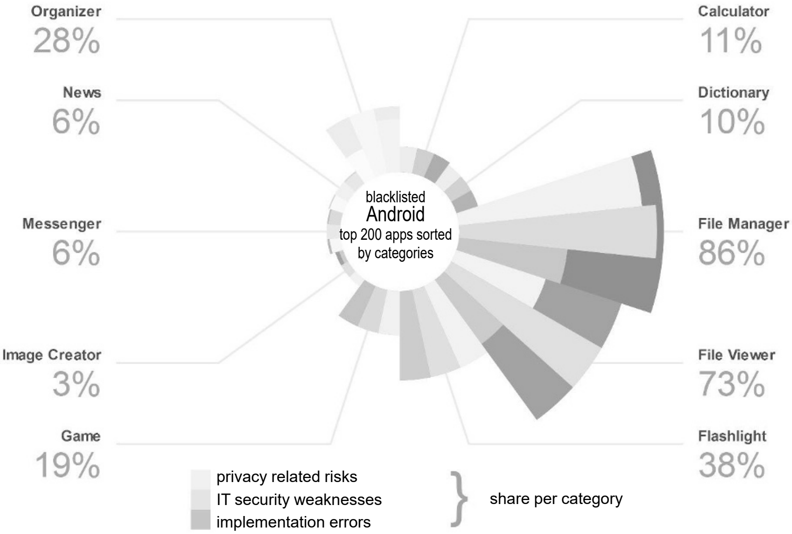


Fig. 16.2 Vulnerabilities of iOS apps by application area according to the Appicaptor Security Index (as an illustration of [6], Fraunhofer SIT)

systems – and an even greater one for the systems of tomorrow. Also increasingly of note are attacks by white-collar criminals, secret service surveillance, and organized crime. Alongside forms of extortion, e.g. by means of so-called ransomware [13], targeted attacks are being conducted on company managers who usually possess multiple authorizations to access sensitive information in their respective companies. Current research findings demonstrate the state of (in-)security of modern information technology at all levels. Examples include:

- *The human factor*: Users are massively overstretched in terms of configuring email client encryption [4]. The following expert comment puts the issue in a nutshell: “In practice, using encrypted e-mail is awkward and annoying” [5]. *Apps*: Three quarters of apps with file access have security issues [6]. *Security apps*: Even apps such as password programs, where the core function is to increase IT security, at times demonstrate serious deficiencies [7]. *Recycled software*: Weak points of freely available software are transferred to a variety of apps by software developers copying and pasting [19], rendering all of these apps vulnerable.

- *Internet infrastructure:* Three quarters of the DNS infrastructure of companies is vulnerable to attack. Two thirds of DNSSEC keys are weak and thus breakable [9].
- *Hardware and embedded software:* Security analyses have revealed numerous embedded software vulnerabilities and have also proven that the encryption implemented via established encryption processes such as RSA can be cracked, if the implementation of the encryption process is susceptible to side-channel attacks [20][21]. Internet of Things (IoT): Security techniques in the Internet of Things are being adapted extremely slowly and many low-end IoT devices have no update or management access at all and thus cannot be patched [10].

Cybersecurity research thus faces enormous challenges that require not only technological innovations, but also a rethinking of the development and operation of secure, software-intensive cyber-physical systems.

16.3 Cybersecurity: relevant for every industry

The specific risks that result from this insecure information technology in products, services, and infrastructures become clear when we look at the various areas of application of information and communications technology and the sectors of industry affected.

Industry 4.0: In the world of production, a fourth industrial revolution is taking place where production systems are being integrated and networked by means of information and communications technology, both vertically with other business systems as well as horizontally across entire value chains. This allows production processes to be designed more flexibly and efficiently and to be adjusted in real time. The openness and interconnectedness of components in these production processes harbors the risk of IT manipulation by attackers, e.g. by competitors or extortionist hackers. If IT vulnerabilities are found in industry control systems this often produces a large number of attack points since products from a small number of manufacturers are generally in use among a large number of users. A summary of the challenges and potential solutions for cybersecurity in Industry 4.0 can be found in [15].

Energy supply: The energy transition relies on decentralization and intelligent consumption management, e.g. in the context of smart grids. To achieve this, devices for energy use and components of energy supply must be linked via information

Fig. 16.3 Digitization opens up opportunities for optimization and new industrial value chains, but also harbors risks due to unauthorized access to sensitive systems and information (Fraunhofer SIT).

technology. This produces entry points for cyberattacks that can lead to economic risks to those affected or even to problems of supply.

Mobility: Mobility is now inconceivable without the use of information technology. Within any one vehicle, the monitoring and control of processes is handled by a number of linked and embedded systems. For traffic management and control purposes, these systems exchange information both between different vehicles as well as between vehicles and infrastructure components. Here, too, there are numerous opportunities for cyberattacks, whereby attackers are able to access vehicles and infrastructure components from a distance via networks, without requiring physical access. Tomorrow's automotive products can thus only be used without risk to life if they are resistant to cyberattacks [17][18].

Finances: The players in the financial world are already highly connected with one another via information and communications technology today. These connections form the nervous system for the economy and industry. Breakdowns and manipulation could lead to huge economic losses. Today a bank robbery no longer requires the use of firearms – all an attacker needs is a computer to access the bank's IT systems. It is thus highly important that financial systems are secure and remain available. This requires cybersecurity.

Logistics: Logistics is the backbone of industry. The monitoring and control of modern logistics processes is today provided in real time via information and communications technology. Transported goods identify themselves electronically. Information technology increases the effectiveness and efficiency of logistics whilst simultaneously reducing vulnerability to human error. Modern logistics must nevertheless be safe from cyberattacks.

Healthcare: Doctors, hospitals, and insurers today rely heavily on information and communications technology, contributing to an increase in efficiency in healthcare and a reduction in costs. Security and data protection requirements are particularly important for medical data.

Public administration: Citizens rightly have high standards regarding IT security with respect to public security, administrative efficiency improvements, safeguarding civil rights and informational self-determination, as well as critical infrastructure provided by the state.

Software: Cybersecurity is becoming ever more important for the software industry. Practically every company utilizes application software in business processes that are critical to their respective business success. This application software is characterized by special functions required for the most diverse range of purposes. Nevertheless, when application software is developed often only the functions relevant to the application domains is considered. Since cybersecurity is only considered marginally in these cases, and due to software's increasing complexity, security loopholes inevitably ensue.

Cloud: For cloud services there are security requirements that go far beyond those for "traditional" application software [23]. Making IT resources available via cloud services has particular economic potential, especially for small and mid-size companies for whom there are significant costs to running in-house IT departments. Cloud computing offers companies the opportunity to reduce investment and operational costs while at the same time increase agility. The high availability requirements for cloud services, combined with their exposed location on the Internet, provide challenges for updating and patching processes and necessitate high to maximum levels of protection from attack, permanent monitoring of the threat landscape, and elaborate attack and intrusion detection mechanisms.

16.4 The growing threat

Today, innovation takes place almost exclusively with the aid of or by means of information technology. The short innovation cycles in the IT industry lead to a constant pressure to modernize, accompanied by an equally constant pressure to reduce costs. Whether it is the Internet of Things, Industry 4.0, big data, blockchain, artificial intelligence, or machine learning: every new trend in information technology intensifies the interaction between information, services, and end devices. This constantly opens up new threats with regard to security and data protection.

16.5 Cybersecurity and privacy protection in the face of changing technology and paradigms

The technological transformation in recent years has strongly changed IT-related security considerations. Traditional IT security was concerned with IT systems, in particular with the protection of IT networks and devices. With the growth in information processing, data security and data protection have increasingly moved to the center of attention. IT security has accordingly increasingly been joined by information security, which is concerned with the protection of information. Here, in contrast to traditional IT security, analog information that is not part of the digital world is also included. With the increasing elimination of the boundaries between the digital and analog worlds promoted by the Internet of Things or the concept of

Security ad hoc,
primarily reactive

»Security & Privacy by Design«
Systematic security, proactive and attack-tolerant

»Cybersecurity at Large«
Systematic security for large real systems, for example „industrie 4.0“ or „smart services“

Fig. 16.4 Paradigm development in cybersecurity: from reactive security, through “security and privacy by design”, to “cybersecurity at large” (Fraunhofer SIT)

Industry 4.0, values distant from IT nevertheless increasingly stand at the center of security technology considerations now.

If cybersecurity is not taken into account across a product's entire lifecycle, then this leads to negative effects for providers: either products and services are not secure enough, or cybersecurity becomes far more expensive than necessary. It is thus important to consider cybersecurity right from the design stage and in the development and integration of new products, services, and production processes.

"Security and privacy by design" is the term used to refer to the consideration of cybersecurity and privacy protection throughout the entire lifecycle of IT products and services [12]. Paying attention to security questions at the earliest possible stage is especially important, since most disclosed weak points are based on errors in the design and implementation of IT systems.

"Security at large" considers security not only during the design and implementation of products and services but also during the integration of IT components for large and complex systems. This also encompasses taking account of the requirements of specific fields of application and technology where numerous components are integrated into large, complex systems. These include, for example, business software, cyber-physical systems, cloud computing, critical infrastructure, or Industry 4.0, and in particular the Internet itself as the fundamental infrastructure in the IT domain.

16.6 Cybersecurity and privacy protection at every level

As the following demonstrates, example solutions are in place at different levels, including the human being, apps, Internet infrastructure, mobile security, hardware and embedded security, IoT security and security monitoring, but also software security and data sovereignty. Nevertheless, there is still a significant need for research into all of these areas, especially with respect to the challenge of security at large, but also with respect to embedded security and security close to hardware, and the tool-assisted development of secure software and services.

Support for users: The Volksverschlüsselung ("people's encryption") initiative launched by Fraunhofer SIT provides a cryptographic key infrastructure that is the prerequisite for end-to-end encryption. Deutsche Telekom operates the solution in a high-security computing center. The focus of Volksverschlüsselung lies in the infrastructure's construction and in the development of a user app that takes charge of key management on the user side and installs the keys in the "right places" in order to overcome configuration hurdles for users. Volksverschlüsselung features

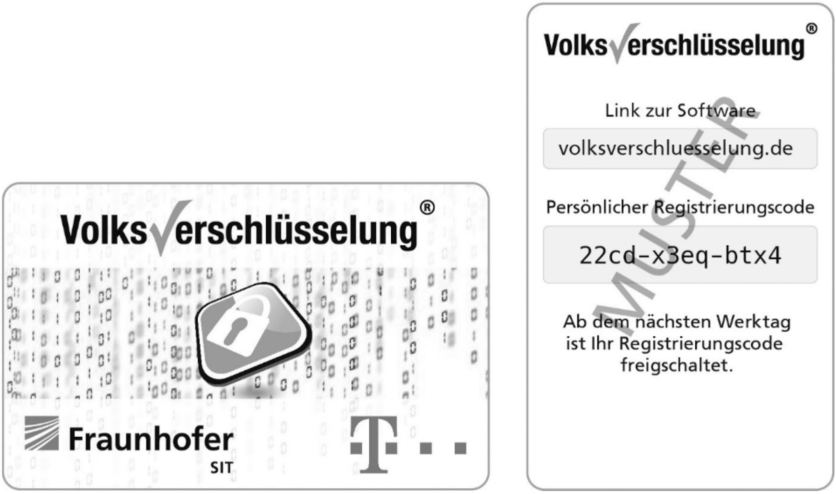


Fig. 16.5 Cryptography is a foundation technology for successful digitization. Fraunhofer SIT, together with Telekom, has introduced Volksverschlüsselung (“people’s encryption”), which is free to use for private individuals. In the context of user registration cards containing registration codes are utilized, for example. (Fraunhofer SIT)

privacy by design as well as usability by design and is structured as a scalable security-at-large solution.

Mobile app: For mobile apps, Fraunhofer SIT and Fraunhofer AISEC have respectively developed a semi-automated service (Appicaptor) and an automated tool (AppRay) for security analysis. Both analysis tools analyze not only Android apps but iOS ones, too. Appicaptor brings together several tools for analysis and test case generation, ranking apps using a range of different non-static techniques. These security tests facilitate fast and practical testing for known errors and implementation weaknesses, amongst others, by means of mass testing where apps are examined quickly and with a minimal failure rate for a specific class of error. Here, researchers gathered characteristics of different error behaviors that correspond to attacks or may be able to be used for attacks so that a practically useful catalog of potential attacks (or error behavior supporting attacks) was produced, enabling to test the resilience of Android and iOS apps to these attacks in various scenarios. This analysis technique also enables testing in cases where the source code of the app is unavailable. Appicaptor detects, amongst others, a range of problems that arise via the incorporation of foreign code in apps and thus represents a significant contribu-

Fig. 16.6 Without end-to-end encryption, emails can be intercepted en route and read like a postcard. (Fraunhofer SIT).

tion to security at large. The AppRay analysis tool enables the flow of data and information in apps to be automatically revealed, along with infringements of data protection guidelines, for example, or of other security rules that are individually configurable.

Enterprise apps: The large codebases of enterprise apps make access for analysis purposes difficult. Here, Fraunhofer SIT introduced the Harvester solution to solve part of the problem of software security analysis in the context of security at large: Harvester extracts relevant runtime values from an app completely automatically, even when obfuscation and anti-analysis techniques such as encryption have been used. The precise program code that calculates the relevant values is cut out from an app and then executed in a secure environment/in isolation. Irrelevant program statements are thus first removed and the amount of code for examination thus minimized.

The cut out code is then executed directly in a separate area (without e.g. waiting for a holding time or restart). In this way, it becomes possible to solve a complex, dynamic analysis problem. Harvester can be used at various points, both as a plugin and also as an independent tool or in-house webservice. Harvester is aimed at diverse user groups such as developers, security experts within companies and security authorities, app store operators, and antivirus providers.

Internet infrastructure: Despite intensive research and standardization activities, essential mechanisms in the Internet continue to be far away from offering sufficient security. One example of this is the naming system in the Internet (domain name system, DNS), in particular the caching strategies used. The topology and architecture of name servers usually utilizes temporarily stored caches. A range of serious vulnerabilities and misconfigurations have been identified: large-scale experiments and measurements show that these caches were generally run very unprofessionally, thus leading to potential for attack and decreased performance [8]. Anyone in a position to manipulate DNS can intercept email and telephone calls or conduct practically undiscoverable phishing attacks and thus gain access to login data and passwords, for example. Fraunhofer SIT is thus working on tools to allow Internet infrastructure to be better secured and is developing recommendations for actions for manufacturers and network operators in order to be able to respond to vulnerabilities, also at a short notice [8].

Hardware and embedded security: Hardware and software protection measures for increasing the security of electronic and digital devices and components have formed part of Fraunhofer AISEC's offering for years now. The institute develops personalized and tailored solutions for different sectors and products. This may be for embedded systems in industrial machinery and plant construction for example, for embedded hardware and software components in industrial control systems and

Fig. 16.7 Without analyses, the dangers involved in using an app cannot be assessed (Fraunhofer SIT)

in the automotive or avionics field, but also for IoT components in home automation or healthcare. Embedded systems are mostly made up of an assembly of several chips and are generally easily physically accessible. They are thus defenseless to attackers with competencies in the fields of electronics, telecommunications, implementations, and hardware attacks. In addition, attackers have the opportunity to exploit internal interfaces such as debug interfaces or to gain direct access to an integrated memory chip. It is thus imperative to strive for a high degree of hardware security for these kinds of systems right from the start. Core areas of focus at the AISEC are in developing secure system-on-chip solutions [24], protecting embedded software from manipulation [25][33], but also in safeguarding secure digital identities for embedded components.

Secure IoT and data sovereignty: Insecure configurations and insufficient monitoring of IoT devices pose a high risk of potential attacks/manipulation, above all in the field of industrial automation/Industry 4.0. If companies use the data from these devices as the basis for decision-making within their business processes then this can have fatal consequences. The Trusted Connector developed at Fraunhofer AISEC protects sensitive business processes from threats that arise due to networking. It ensures that only reliable data is incorporated into critical decision-making. The pre-processing of this data by applications in the Trusted Connector facilitates reliable processing chains within the company and beyond company borders. A secure container-based execution environment facilitates strict isolation of running applications. Data and software are thus protected from loss and unwanted modification. Integrity verification for the data and installed applications combined with a hardware-based security module (Trusted Platform Module – TPM) together ensure a high level of reliability [25]. This is supplemented with flexible monitoring of access and dataflows facilitating fine-grained organization of dataflows both within and outside of the company [26]. The AISEC Trusted Connector is also the central security component in the Industrial Data Space (IDS) that is currently being developed by Fraunhofer together with partners from industry [34]. The IDS aims to create a reference architecture for a secure data space that enables companies from various industries to manage their data assets confidently. This data space is based on a decentralized architectural approach where data owners are not required to surrender their data superiority or sovereignty. A central component of the architecture is the Industrial Data Space Trusted Connector which facilitates the supervised exchange of data between participants in the Industrial Data Space.

Continuous security monitoring and security assessment: Today's IT-based systems are largely dynamic: new components are added during operation, communications partners change, but also new software artifacts such as apps and software updates are loaded and executed while operation is ongoing. Techniques are being developed at Fraunhofer AISEC to continuously evaluate the current security state of IT-based systems such as cloud platforms [27]. Using advanced analysis techniques based on machine learning, malicious code can be identified early, for example, so that potential damage can be minimized [28]. Processes developed specially at AISEC also enable security analysis to be carried out via encrypted communications paths [31] such that systems can be supervised from afar, for example, without losing the protection of the secure communications channel. Using the isolation and supervision techniques developed at AISEC [30] as well as measures for continuous integrity measurement [29], a system can continuously compare its system state to specified rules and requirements to be observed, and identify and defend against deviations early on that indicate potential preparatory steps for attacks.

Software security: Cyber-physical systems are software-intensive systems that are operated as original systems with new innovations integrated. At Fraunhofer AISEC, software development methods and tools that cover the entire lifecycle of software artifacts are being researched. To achieve this, constructive measures are being developed in order to plan for security right from the design stage and to take appropriate account of it during integration and configuration [18][22]. On top of that, software tools are being developed to analyze software for potential weak points before it is commissioned, with as high a degree of automation as possible, and to overcome these weak points as far as possible, automatically and without semantic alteration [32]. Using the encapsulation techniques provided such as isolating containers, insecure third-party/legacy system components that cannot be hardened can also be securely integrated into complex value creation networks, such that interaction between secure and insecure components is possible while demonstrably maintaining the required security characteristics.

Martin Priester, Fraunhofer Academy

Securely ready for digitization

Lifelong learning means being able to keep up. New knowledge enables limits to be pushed back and solutions to be found to new kinds of problems where the old approaches are no longer promising. A quick look at the rapid developments in the field of IT security in recent years provides a realistic idea of what "keeping up" means: more than 300,000 new variants of malicious software are discovered every day, according to Bitkom [35]. The Internet not only enables cybercriminals to digitize traditional offenses such as fraud, extortion, and vandalism, but also to establish new business models as shown by the botnet avalanche [36]. There is no sign that the pace of development in the IT security field will lessen, but there are nevertheless indications that some actors – whilst trying to keep up – are running out of steam.

Studies on the shortage of specialist IT security staff paint a frightening picture that can rightly be described as a war for talent [37]. Clearly, the demand for specialist staff cannot be sufficiently met by the number of university graduates and vocationally trained IT professionals. Increasing digitization additionally presents the staff of companies and authorities with new challenges in terms of qualification. This is because IT security cannot be safeguarded purely by those commissioned for it. This means that the diversity of attack vectors must be reflected in protective and qualification measures that incorporate large parts of the staff.

A quick look at the instruction by the Federal Office for Information Security [38] shows how extensive the security specific competency profile is. According to the document, the typical weak points that make IT systems vulnerable can be divided into four categories:

1. Software programming errors
2. Weak points in the software's specification
3. Configuration errors
4. Users of IT systems as an insecurity factor

We need only remember that the source code in commonly used software products may be several million lines long to understand how great the task of developing secure software is. Security by design approaches only find their way into development praxis, however, if the knowledge of, for example, security-oriented programming languages forms part of vocational or practical training courses. At the same time, the application of principles for secure software development also requires a response to organizational questions. How should the interfaces between the different roles (developer, tester, system integrator, etc.) be best designed, for example?

However, resisting threats is just one of the elements. It is equally important that companies and authorities identify successful attacks in the first place. They need to be able to assess the degree of damage and find ways to "get affected systems going" again. This knowledge is spread across several individuals within the company and can only be retrieved through all of the different staff working well together.

But how can companies and authorities be in a position to carry on keeping up. How can IT security risks be reduced and better overcome? Professional development is the decisive key here. Thereby four conditions need to be met so that new knowledge of IT security can quickly be brought into effect.

1. Only the latest research knowledge allows organizations to be a step ahead of potential attackers and to use, for example, new development and testing processes for secure software.
2. Not all actors have the same need for knowledge. Professional development content must be tailored for specific topics, roles, and sectors and take account of differing levels of prior knowledge.
3. Knowing comes from doing. Training in laboratories with current IT infrastructure allows participants to experience actual threat scenarios and test out the applicability of solutions.
4. Utilize limited time resources sensibly. Compact formats with practical issues at the forefront can be meaningfully integrated into day-to-day professional life and combined with appropriately designed learning media made available electronically.

With the Cybersecurity Training Lab initiative, Fraunhofer is developing a professional development program that meets these challenges, together with selected colleges. It combines the expertise of the different partners in the various fields of IT security applications and activities (such as, for example, software and hardware development, IT forensics, and emergency response) into a robust association for research and qualification.

Sources and literature

[1] https://www.bundesdruckerei.de/de/studie-it-sicherheit, Abruf am 11.7.2017

[2] https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/kompetenzen-fuer-eine-digitale-souveraenitaet.pdf? blob=publicationFile&v=14, Abruf am 11.7.2017

[3] https://www.bitkom.org/Presse/Presseinformation/IT-Sicherheit-Cloud-Computing-und-Internet-of-Things-sind-Top-Themen-des-Jahres-in-der-Digitalwirtschaft.html, Abruf am 11.7.2017

[4] Fry, A., Chiasson, S., & Somayaji, A. (2012, June). Not sealed but delivered: The (un) usability of s/mime today. In Annual Symposium on Information Assurance and Secure Knowledge Management (ASIA'12), Albany, NY.

[5] https://arstechnica.com/security/2013/06/encrypted-e-mail-how-much-annoyance-will-you-tolerate-to-keep-the-nsa-away/3/ , Abruf am 21.7.2017

[6] https://www.sit.fraunhofer.de/de/securityindex2016/, Abruf am 12.7.2017

[7] https://codeinspect.sit.fraunhofer.de, Abruf am 13.7.2017

[8] Klein, A., Shulman, H., Waidner, M.: Internet-Wide Study of DNS Cache Injections, IEEE International Conference on Computer Communications (INFOCOM), Atlanta, GA, USA, May 2017.

[9] Shulman H., Waidner M.: One Key to Sign Them All Considered Vulnerable: Evaluation of DNSSEC in Signed Domains, The 14th USENIX Symposium on Networked System-sDesign and Implementation (NSDI), Boston, MA, USA, March 2017.

[10] Simpson, A. K., Roesner, F., & Kohno, T. (2017, March). Securing vulnerable home iot devices with an in-hub security manager. In Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on (pp. 551-556). IEEE, 2017

[11] Solms, R. and Van Niekerk, J., 2013. From information security to cyber security. computers & security, 38, pp.97-102. 2013

[12] Waidner, M., Backes, M., Müller-Quade, J., Bodden, E., Schneider, M., Kreutzer, M., Mezini, M., Hammer, Chr., Zeller, A. Achenbach, D., Huber, M., Kraschewski, D.: Entwicklung sicherer Software durch Security by Design,. SIT Technical Report SIT-TR-2013-01, Fraunhofer Verlag, ISBN 978-3-8396-0567-7, 2013

[13] Claudia Eckert: IT-Sicherheit: Konzepte – Verfahren – Protokolle, 9th Edition, De Gruyter, 2014

[14] Claudia Eckert. „Cybersicherheit beyond 2020! Herausforderungen für die IT-Sicherheitsforschung“. In: Informatik Spektrum 40.2 (2017), pp. 141–146.

[15] Claudia Eckert. „Cyber-Sicherheit in Industrie 4.0“. In: Handbuch Industrie 4.0: Geschäftsmodelle, Prozesse, Technik. Ed. by Gunther Reinhart. München: Carl Hanser Verlag, 2017, pp. 111–135.

[16] Bundesamt für Sicherheit in der Informationstechnik (BSI), „Die Lage der IT-Sicherheit in Deutschland 2016“, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/

[17] Martin Salfer and Claudia Eckert. „Attack Surface and Vulnerability Assessment of Automotive Electronic Control Units“. In: Proceedings of the 12th International Conference on Security and Cryptography (SECRYPT 2015). Colmar, France, July 2015.

[18] D. Angermeier and J. Eichler. „Risk-driven Security Engineering in the Automotive Domain“. Embedded Security in Cars (escar USA), 2016.

[19] F. Fischer, K. Böttinger, H. Xiao, Y. Acar, M. Backes, S. Fahl, C. Stransky. „Stack Overflow Considered Harmful? The Impact of Copy & Paste on Android Application Security" , IEEE Symposium on Security and Privacy 2017.

[20] A. Zankl, J. Heyszl, G. Sigl, „Automated Detection of Instruction Cache Leaks in RSA Software Implementations", 15th International Conference on Smart Card Research and Advanced Applications (CARDIS 2016)

[21] N. Jacob, J. Heyszl, A. Zankl, C. Rolfes, G. Sigl, „How to Break Secure Boot on FPGA SoCs through Malicious Hardware", Conference on Cryptographic Hardware and Embedded Systems (CHES 2017)

[22] C. Teichmann, S. Renatus and J. Eichler. „Agile Threat Assessment and Mitigation: An Approach for Method Selection and Tailoring". International Journal of Secure Software Engineering (IJSSE), 7 (1), 2016.

[23] Niels Fallenbeck and Claudia Eckert. „IT-Sicherheit und Cloud Computing". In: Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung, Technologien, Migration", ed. by Thomas Bauernhansl, Michael ten Hompel, and Birgit Vogel-Heuser. Springer Vieweg, 2014, pp. 397–431.

[24] N. Jacob, J. Wittmann, J. Heyszl, R. Hesselbarth, F. Wilde, M. Pehl, G. Sigl, K. Fisher: „Securing FPGA SoC Configurations Independent of Their Manufacturers", 30th IEEE International System-on-Chip Conference (SOCC 2017)

[25] M. Huber, J. Horsch, M. Velten, M. Weiß and S. Wessel. „A Secure Architecture for Operating System-Level Virtualization on Mobile Devices". In: 11th International Conference on Information Security and Cryptology Inscrypt 2015.2015.

[26] J. Schütte and G. Brost. „A Data Usage Control System using Dynamic Taint Tracking". In: Proceedings of the International Conference on Advanced Information Network and Applications (AINA), March 2016.

[27] P. Stephanow, K. Khajehmoogahi, „Towards continuous security certification of SoftwareasaService applications using web application testing", 31th International Conference on Advanced Information Networking and Applications (AINA 2017)

[28] Kolosnjaji, Bojan, Apostolis Zarras, George Webster, and Claudia Eckert. Deep Learning for Classification of Malware System Call Sequences. In 29th Australasian Joint Conference on Artificial Intelligence (AI), December 2016.

[29] Steffen Wagner and Claudia Eckert. „Policy-Based Implicit Attestation for Microkernel-Based Virtualized Systems". In: Information Security: 19th International Conference, ISC 2016,Springer 2016, pp. 305–322.

[30] Lengyel, Tamas, Thomas Kittel, and Claudia Eckert. Virtual Machine Introspection with Xen on ARM. In 2nd Workshop on Security in highly connected IT systems (SHCIS), September 2015.

[31] Kilic, Fatih, Benedikt Geßele, and Hasan Ibne Akram. Security Testing over Encrypted Channels on the ARM Platform. In Proceedings of the 12th International Conference on Internet Monitoring and Protection (ICIMP 2017), 2017.

[32] Muntean, Paul, Vasantha Kommanapalli, Andreas Ibing, and Claudia Eckert. Automated Generation of Buffer Overflows Quick Fixes using Symbolic Execution and SMT. In International Conference on Computer Safety, Reliability & Security (SAFECOMP), Delft, The Netherlands, September 2015. Springer LNCS.

[33] M. Huber, J. Horsch, J. Ali, S. Wessel, „Freeze & Crypt: Linux Kernel Support for Main Memory Encryption" ,14th International Conference on Security and Cryptography (SECRYPT 2017).

[34] B. Otto et. al: Industrial Data Space, Whitepaper, https://www.fraunhofer.de/de/forschung/fraunhofer-initiativen/industrial-data-space.htm

[35] https://www.bitkom.org/Presse/Presseinformation/Die-zehn-groessten-Gefahren-im-Internet.html Abruf am 30.06.2017

[36] L. Heiny (2017): Die Jagd auf Avalanche. http://www.stern.de/digital/online/cyberkriminalitaet--die-jagd-auf-avalanche-7338648.html Abruf am 30.06.2017

[37] M. Suby, F. Dickson (2015): The 2015 (ISC)^2 Global Information Security Workforce Study. A Frost & Sullivan White Paper.

[38] https://www.allianz-fuer-cybersicherheit.de/ACS/DE/_/downloads/BSI-CS_037.pdf? blob=publicationFile&v=2 Abruf am 30.06.2017

Fault-Tolerant Systems

17

Resilience as a security concept in the era of digitization

Prof. Dr. Stefan Hiermaier · Benjamin Scharte
Fraunhofer Institute for High-Speed Dynamics,
Ernst-Mach-Institut, EMI

Summary

The more we become dependent on the functioning of complex technical systems, the more important their resilience becomes: they need to maintain the required system performance even when internal and external failures and disruptions occur. This applies both to individual systems (e.g. cars, medical devices, airplanes) as well as to infrastructure (traffic, supply systems, information and communications systems). Designing these complex systems to be resilient requires Resilience Engineering, that is, a process of maintaining critical functions, ensuring a graceful degradation (in the case where the critical functionality cannot be retained due to the severity of the disruption) and supporting the fast recovery of complex systems. This necessitates generic capabilities as well as adaptable and tailored technical solutions that protect the system in the case of critical issues and unexpected or previously nonexistent events. Cascade effects that occur in critical infrastructures during disruption, for example, may thus be simulated and their effects proactively minimized.

17.1 Introduction

Technical systems that function safely and reliably are vital for our society. 250 years after the beginning of the Industrial Revolution, more than 70 years after the advent of the first computer, and almost 30 years after the invention of the World

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_17

Wide Web, there is hardly any field conceivable without technical systems. In essence, these systems determine the day-to-day lives of everyone living in modern industrial societies. This triumph is due to the fact that the systems make day-to-day lives easier in a multitude of ways. In industry, they provide efficiency and quality gains, resulting in better products and services. For recreational life, they make activities possible that were previously either impossible or at least required significant expense, from long-distance travel to e-sports.

The increasing digitization of industry, work, and private life is an additional and, in view of its far-reaching effects, even revolutionary step towards a world that is completely dependent on technical systems. The various chapters of the present book span themes such as additive manufacturing, digital factories, and individualized mass production, from the question of the usefulness of artificial intelligence for challenging situations, to e-government – digital citizen-oriented administration. The authors highlight opportunities and possibilities arising from these various developments in digitization. But they also shed light upon the specific challenges related to the topics mentioned.

If we take a top-level view of the trends in different fields, namely from a systems perspective, it becomes clear that there is a salient commonality: the success of these developments depends on the ever increasing connectedness of previously separate societal and technical fields. Irrespective of the various positive aspects that increasing connectedness can entail, this also gives rise to challenges at the systems level. Even individual separate systems are becoming increasingly complicated due to their inherent intelligence and are well past the point of anyone but specialists understanding them. Even specialists are starting to reach their limits however. If several complicated systems are connected together, then, with increasing frequency, more complex (technical) systems result.

It is important then, as we progress through digitization, to ensure that these complex technical systems demonstrate the maximum possible fault tolerance both during day-to-day operation, but also and above all in exceptional cases (in the case of disruptions of any kind). Complexity in itself is neutral at first, neither good nor bad. The same is true of interconnecting different systems. In a crisis, however, complexity and connectedness may increase negative effects or give rise to outcomes that were neither foreseen nor planned. For this reason, the traditional security approach of classic scenario-based risk management is no longer sufficient. An enhanced systemic concept is required with which to analyze, understand, and ultimately, increase the fault tolerance of complex technical systems [13]. The present article introduces the need for this kind of concept, the ideas behind it, and a concrete implementation of it in four sections. First, the challenges facing complex technical systems are explained in greater detail. Next, the concept of resilience is presented as a means of dealing effectively with these challenges. Building upon this, the third section takes a look at a specific applied project concerned with valid simulations of cascade effects in complex, coupled network infrastructures and developing measures to improve these kinds of network structures. Finally, the results are summarized and an outlook provided.

17.2 Challenges for fault-tolerant systems

The whole is greater than the sum of its parts. This is precisely the effect we observe when dealing with complexity. Complicated systems consist of a variety of individual parts which are logically connected with one another. Providing an adequate description of their behavior may be extremely difficult in some conditions. They can be explained reductively, however. That is, the behavior of the system as a whole can be deterministically identified by observing the causal relationships of the individual parts of the system to one another. Thus, the computing power available is essentially the only factor in determining whether and how quickly system behavior can be correctly predicted. A complex system, in contrast, cannot be explained on the basis of its individual parts; it is simply more than the sum of its parts. Complex systems are able to develop emergent properties, that is, properties that can only be explained when observing the system holistically [8]. The boundaries between complicated and complex systems are fluid and it is often difficult or even impossible to decide whether an actual system can better be described properly using the one or the other term. As previously mentioned, complexity makes it impossible to understand systems using reductive tools, which represents a challenge to the fault tolerance of these kinds of systems. Traditional risk analysis and risk management make use of explicitly reductive principles ranging from clearly specified scenarios through precisely defined probabilities to exact assessments of damages [15][20]. This is no longer sufficient for analyzing complex technical systems, which is why there is an increasing reliance upon the concept of resilience.

A second challenge for fault-tolerant systems is susceptibility to cascade effects resulting from the ever-increasing networking of different technical systems. The term “cascade” usually refers to a sequence of levels or processes in the context of specific events. A common albeit somewhat inaccurate example is the domino effect. When applied to disruptive events, cascades refer to scenarios where an initial, primary disturbance event is followed by a series of secondary events, which can themselves be viewed as a new disturbance event [17]. A dramatic example of cascading effects is provided by the earthquake off the coast of the Tōhoku region of Japan on March 11, 2011. Measuring 9.0 on the Richter scale, it remains the strong-

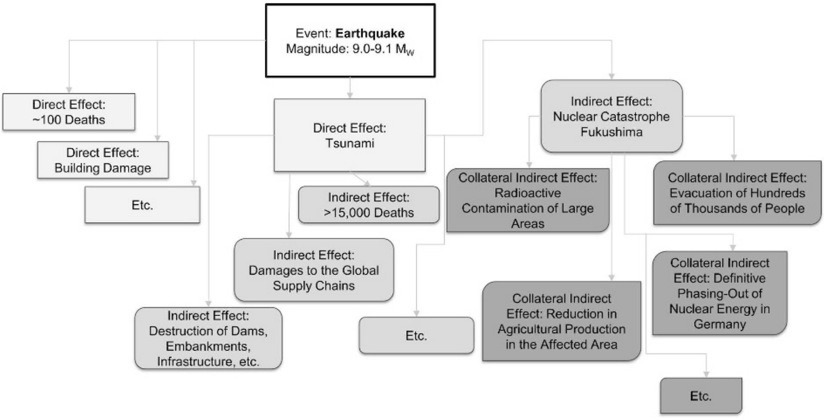


Fig. 17.1 A typical example of serious cascade effects: the Tōhoku earthquake of March 11, 2011 (Fraunhofer EMI)

est quake in Japan since records began and one of the strongest earthquakes ever. Fig. 17.1 shows the direct and indirect (cascading) effects caused by the earthquake. It becomes clear that the majority of the damage was caused by a secondary event caused by the earthquake itself: the tsunami. It was this tsunami that, in a further cascade, led to the Fukushima nuclear catastrophe and, at least indirectly, forced the German federal government at the time to change course towards a definitive phase-out from atomic energy.

The developments illustrated above lead to increased vulnerability of complex technical systems with respect to serious disruptive events. At the same time, the number of these events is also rising. Climate change, for example, is leading to increasingly extreme weather events. Terrorist attacks, too, are becoming more frequent, using different methods to target completely different aims. Cyberattacks on important computer networks and infrastructures are of particular relevance in the context of digitization, of course, such as the WannaCry malware attack in May 2017 (see Fig. 17.2). Over 300,000 computers in more than 150 countries were affected by this attack – the worst yet known – which was conducted with the aid of ransomware, malicious software that effectively “kidnaps” affected computers, demanding payment of specific sums for the release of the encrypted data. In Great Britain, the software affected numerous hospitals, in some cases causing patients to be turned away, ambulances to be diverted, and routine operations to be cancelled. In Germany, advertising billboards and ticket machines belonging to the Deutsche

Fig. 17.2 Screenshot of a computer infected with the WannaCry malware (Fraunhofer EMI)

Bahn (German Railway) were affected, with video surveillance technology in train stations also partially disrupted. And in France, the Renault automotive group had to temporarily suspend production in a number of factories [7][21]. These impacts on critical infrastructures and important areas of industry show how the networking and coupling of different complex technical systems can increase vulnerability thereof in the presence of disruptive events such as cyberattacks.

17.3 Resilience as a security concept for the connected world

In view of the challenges and developments just described, it is extremely clear that our society needs an adequate security concept to prevent the failure of critical systems. Where systems do fail in spite of all efforts, further mechanisms need to be in place that ensure the speediest possible recovery of the relevant functionalities. In order to arm complex connected systems against disturbances that are external as well as internal, expected as well as unexpected, and which occur abruptly as well as those which develop more slowly, a holistic view of system security is needed. The discussion among the security research community, pertaining to how this kind of systemic approach should look and how concrete solutions to increase the fault tolerance of complex connected systems can thus be developed is centered around the terms "resilience" and "resilience engineering". "Resilience" in particular has become a dominant term in security research in recent years.

Pertinent aspects of the term "resilience"

In disciplines such as ecology or psychology, work with the concept of resilience has been ongoing for decades now [5][12][18]. The word itself is of Latin origin: *resilire* means "to spring back". The dictionary defines resilience as "the capacity to recover quickly from difficulties; toughness" [3]. This definition stems from what is probably still the most prominent use of the concept: in psychology, people are described as resilient if they are able to successfully withstand crises. These kinds of crises may include events such as serious illness, the loss of a family member, unemployment, or a difficult childhood overshadowed by poverty, violence, or abuse.

Fig. 17.3 The roly-poly doll, often used to explain resilience, provides an inadequate illustration of the concept (© Beni Dauth, released to public domain)

The ability to successfully overcome crises by means of specific protective factors is not only interesting for people as individuals. Entire societies and their relevant subgroups and subsystems should also possess this faculty. Canadian ecologist C. S. Holling was the first to explore the meaning of resilience with respect to complex systems. Although his work was focused on ecological systems, his reflections and ideas are also relevant for the resilience of technical systems. Just like ecosystems, technical systems fulfill specific functions. They are robust up to a certain point when faced with more limited disruptions, maintaining a stable equilibrium. What threatens the survival of ecosystems above all, according to Holling, are abrupt, radical, and irreversible changes caused by unusual, unexpected, and surprising events. Non-resilient systems conceived only for stability are unable to respond flexibly to these kinds of events due to the deterministic factors that previously facilitated maintenance of the equilibrium, and thus they founder [9][12]. The ability to once again find a (new) state of equilibrium following these kinds of disruptions, a state where the relevant system functions can still/again be provided, is described as resilience.

Resilience can thus be understood as the ability of complex technical systems to successfully overcome crises, namely also when these crises are caused by unexpected, surprising, and serious events. In doing so, the system does not necessarily return to its original state – in this sense, the frequently used image of the roly-poly doll is not really a fitting description of resilience (see Fig. 17.3) – but instead, it is just as likely to achieve a new, stable equilibrium.

A definition of resilience

In his book, *Resilient Nation*, from 2009, Charlie Edwards drew extensively on classic disaster management cycles to provide a better understanding of the far-reaching concept of resilience and give it concrete expression [4]. The resulting resilience cycle may be slightly expanded to consist of the five phases of prepare, prevent, protect, respond, and recover (see Fig. 17.4). The first step is to seriously *prepare* for adverse events, especially in terms of early warning systems. Reducing the underlying risk factors should then *prevent* the occurrence of the event itself as far as possible. If it nevertheless does occur, it is important that physical and virtual systems *protect* from and minimize the negative impacts function without error. In addition, fast, well-organized and effective emergency assistance is required. During this time, the system must be able to maintain its essential ability to function as much as possible (*respond*). After the immediate period of damage has ended, it is important that the system is in a position to *recover* and draw appropriate learning from what has happened so that it is better equipped for future threats [22]. Based

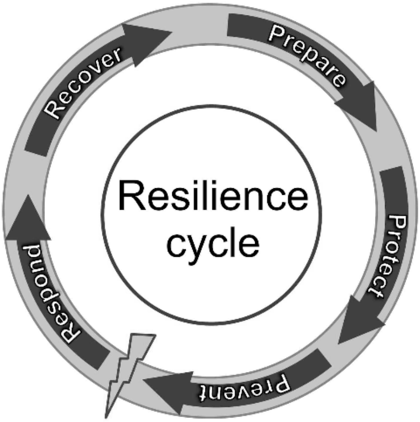


Fig. 17.4 The resilience cycle (Fraunhofer EMI)

on this resilience cycle and the aforementioned aspects, resilience overall can be defined as follows:

> *"Resilience is the ability to repel, prepare for, take into account, absorb, recover from and adapt ever more successfully to actual or potential adverse events. Those events are either catastrophes or processes of change with catastrophic outcome which can have human, technical or natural causes."* [19]

Developing resilient complex technical systems

A particular feature of resilient systems is that they are capable of dynamically responding to constantly changing environmental influences and adapting to unexpected serious events. In this sense, resilience is not a static state but a property of active, adaptive systems which are capable of learning. Resilience has thus grown far beyond its original Latin meaning. Similarly, it is clearly distinct from its understanding in physics and material science where resilience is defined as the ability of a material to be deformed elastically through the influence of energy. Resilience here is measured as the maximum energy that the material can absorb per unit volume without plastic deformation [12][18]. If we "translate" this meaning to complex technical systems, this would imply a pure "bouncing back" to the status quo ante [24]. This term "bouncing back" has had an astounding career in discussions surrounding resilience; engineering science approaches in particular have tried to use this catchy description to give meaning to resilience.

Returning to an initial state of whatever kind subsequent to a disturbance is impossible in pure logical terms due to the dynamic environment that complex technical systems exist in, plus their interactions with said environment. Notwithstanding this, the engineering sciences are investigating how the resilience of complex technical systems to disturbances can be developed and their fault tolerance increased. In scientific circles, the term "Resilience Engineering" is currently gaining prominence as a suitable term for describing the process of increasing resilience with the aid of engineering solutions [2][23]. The term was coined in recent years by researchers such as Erik Hollnagel and David Woods. According to them, the focus of measures to increase security traditionally lay on protective concepts for common threat scenarios. This is where Resilience Engineering, as understood by Hollnagel and Woods, comes in. It is a question of including the possible and not merely the probable in planning, implementation, and execution. New and unexpected threats in particular, since they may differ in extent from all of the scenarios considered, provide systems with challenges that are able to be met with the aid of Resilience Engineering [14][16][25]. Systems of all kinds need to demonstrate sufficiently large and appropriate security margins [14]. This orientation around the possible rather than the probable leads to a necessity for reorientation, namely away from the damaged and towards the normal state. After all, everyday life shows that things normally function as they should. It is unusual for something to go (seriously) wrong. Even complex systems operate relatively smoothly under normal circumstances. Understanding the functioning of complex systems is the appropriate and necessary requirement for identifying and minimizing potential faults, problems, and risks for these systems [11]. This understanding of Resilience Engineering teaches us several things about how complex systems can be designed to be resilient. Nevertheless, Hollnagel, Woods, and their colleagues focused less on complex technical systems and more on organizationally complex systems such as hospitals or air traffic control. Thus, their ideas for Resilience Engineering need to be further developed if they are to make a contribution to the fault tolerance of complex technical systems.

Resilience Engineering – a definition from engineering science resilience research

The first priority is to maintain the critical functionality of the system in question as much as possible, even in exceptional cases. As already mentioned on numerous occasions, complex technical systems always serve a defined purpose, such as supplying society with energy. If a potentially catastrophic event occurs, the system itself may very well be completely changed or severely damaged. The determining factor in the event of damage is to maintain critical subfunctions of the system in a controlled manner, even beyond standard requirements, and thereby avoid a catastrophic total breakdown [2]. Here, we see a partial reflection of Hollings' idea of the different states of equilibrium that ensure a system's survival. In the case that critical functionality cannot be maintained due to the severity of the disruptive event, "graceful degradation" must at least be ensured. That is, an abrupt collapse of the entire functionality must be avoided, providing the system's operators and rescue services with sufficient time to make functional alternatives available. As soon as the event is over, technical systems developed using Resilience Engineering begin to recover from the effects. This fast recovery from damage does not only incorporate bouncing back to the original state but also the implementation of learning drawn from the experience, and an adaptation to changed circumstances [23]. A key component of Resilience Engineering is providing complex technical systems with generic capabilities. This idea was borrowed from the concept of generic competencies that allow people to successfully overcome adverse events, even unexpected or hitherto nonexistent events. For example, an experimental study was able to demonstrate that exercising generic competencies, compared with strictly keeping to precisely stipulated rules and procedures, can increase the success of a ship's crew when dealing with critical situations [1]. Unlike people, however, technical systems do not possess the ability a priori to improvise or adapt flexibly to changing situations. This requires Resilience Engineering to implement generic capabilities as heuristics in technical systems. Examples of these kinds of heuristics are redundancy, the availability of backups, foreseeability, complexity reduction, and other properties [10][14]. It may for example specifically be a question of accelerating research towards new methods for modeling and simulating complex systems that are able to simulate and investigate the impacts of adverse events, in particular with respect to cascade effects (see Ch. 17.4).

At the same time, however, Resilience Engineering also signifies the targeted use of the latest innovative technologies for the design and utilization of complex technical systems. These technologies need to be customized for specific systems and specific tasks. The efficient use of customized technologies to optimize the functionality of complex technical systems during normal operation is one option for increasing the number of processes functioning seamlessly and thus pursuing the kind of Resilience Engineering conceived by Hollnagel and Woods. Overall, Resilience Engineering offers complex technical systems the opportunity to successfully interact with both known problems (by means of customized technologies) as well as with unexpected interruptions or even previously nonexistent crises (thanks to generic capabilities). In summary, the concept can be defined as follows:

> *"Resilience Engineering means preserving critical functionality, ensuring graceful degradation and enabling fast recovery of complex systems with the help of engineered generic capabilities as well as customized technological*

solutions when the systems witness problems, unexpected disruptions or unexampled events.” [23]

17.4 Applied resilience research: Designing complex connected infrastructures that are fault-tolerant

In the previous sections, the terms “resilience” and “Resilience Engineering” were defined, and we explained why, in view of existing developments and challenges, the fault tolerance of complex technical systems can only be raised by making use of these kinds of holistic concepts. Building on these ideas, the following section introduces a specific application project that allows for the simulation and understanding of cascade effects in complex coupled network infrastructures. A tool for designing and analyzing resilient technical systems should ideally possess a range of capabilities. It must for example be able to model the physical components of the system and their interactions, define target system performance, and compare actual and target performance. The opportunity to feed load cases caused by specific

☑ Damage Probability ☑ Damage Propagation ☑ Cascading Effects ☑ Vulnerability ☑ Mitigation Strategies

Main Slider Time Step: None

Fig. 17.5 Illustration of a system of coupled network infrastructures in Helsinki in their undisturbed state (Fraunhofer EMI)

Fig. 17.6 Effects of a storm on the system of coupled network infrastructures in Helsinki (Fraunhofer EMI)

disruptive events into the system as well as generic (that is, event-independent) damage scenarios and to simulate their effects is also important. In doing so, it should be possible to facilitate both identification of critical system components as well as assessment of the fault tolerance of the system. Subsequently, measures to increase resilience can be integrated and, with the aid of a new calculation, the resilience of the improved system evaluated and compared with that of the original system. CaESAR[1] is a software tool developed at Fraunhofer EMI for simulating and analyzing coupled network infrastructures that demonstrate a large number of these abilities.

CaESAR is designed to simulate cascading effects within and especially between various coupled infrastructures. The first systems considered here are the energy grid, water supply, and mobile telephony network. These networks are shown on an overview dashboard as nodes and arcs on a georeferenced map. Fig. 17.5 shows an example of the networks identified in the Finnish capital of Helsinki. CaESAR includes a “crisis editor” which is used to either implement specific disruptive events

---

[1] Cascading Effect Simulation in Urban Areas to assess and increase Resilience

based on actual threat scenarios such as a storm of strength X, or otherwise to populate generic damage scenarios. These disruptive events may occur individually or in combination with one another. In addition, the events can be allocated different intensities and a definitive chronological sequence in the editor. Fig. 17.6 illustrates the effects of a storm on Helsinki's various network infrastructures, for example.

In the next step, CaESAR uses a flow model to simulate how the disruptive events spread through the various coupled networks. Here, the software includes interfaces for tools capable of simulating damage propagation in greater detail within individual networks such as the power grid. The damage to the overall system of coupled networks is determined via sensitivity analysis in order to calculate the probability of failure of individual components and known failure mechanisms. The result is a residual performance level for the system after the disruption. In order to identify critical components and failure mechanisms, the probabilities in the sensitivity analysis are gradually varied. Criticality here means that the components either fail very often and/or that their failure causes particularly extensive (cascading) damage within the system as a whole. The data thus produced is used to provide a resilience score for the system. CaESAR is simultaneously able to suggest measures to overcome the weak points identified. To this end, a package of predefined measures is currently integrated in the software from which the user is able to make appropriate selections and analyze their effects on the system's loss of performance with respect to one or several disruptive events.

In summary, CaESAR makes it possible to simulate complex technical systems (in this case, coupled network infrastructures) and their behavior in the face of different damage scenarios, including generic ones. This represents an important step towards increasing the fault tolerance of these kinds of systems as part of Resilience Engineering. The aim is to enhance CaESAR over the medium term and equip it to additionally simulate other infrastructure systems and their shared connections as well as the effects of various disruptive events on these systems. In order to identify the challenges of networking societally relevant systems, over the course of digitization, new approaches and tools similar to CaESAR should be developed in future.

17.5 Outlook

When the whole really is more than the sum of its parts – and in view of the various complex systems which exist in our everyday lives, there is no doubt that this is the case – we need a systematic view from above or from outside in order to understand "the whole". Due to the ongoing digitization of our society, more and more previously separate fields are being connected with one another. In order to be able to nevertheless guarantee the maximum possible fault tolerance of societally relevant systems, a holistic security concept such as resilience is required. This may be implemented with the help of a Resilience Engineering approach, giving rise to tools such as the CaESAR software.

Nevertheless, engineering and technological implementation of resilience principles is still in its relative infancy [2][23]. Many opportunities are currently available here to implement resilience right from the start during the development of new technologies and in particular when they are widely used. One example would be autonomous driving, where questions regarding the security and reliability of systems in many ways play a decisive role. Another would be the digital management of a society's critical infrastructures. Here, too, it is important to ensure that security aspects are taken seriously and integrated into systems as a matter of course in the face of increasing automation and networking. At the same time, however, any resulting potential risks need to be weighed up carefully. Here, too, the concept of resilience provides excellent opportunities with its holistic approach. Analysis of the key themes here is provided by the article in this volume on data security as a prerequisite for digitization.

In summary, we can see that in the area of the resilience of complex technical systems, a range of open (research) questions remain that in future need to be more deeply engaged with academically by both engineering and technology as well as by the natural and social sciences. Intensive work is already being carried out by Fraunhofer-Gesellschaft and Fraunhofer EMI in particular, on innovative solutions to increase the fault tolerance of complex technical systems.

Sources and literature

[1] Bergström J, Dahlström N, van Winsen R, Lützhöft M, Dekker S, Nyce J (2009): Rule- and role retreat: An empirical study of procedures and resilience. In: Journal of Maritime Studies 6:1, S. 75–90

[2] Bruno M (2015): A Foresight Review of Resilience Engineering. Designing for the Expected and Unexpected. A consultation document. Lloyd's Register Foundation, London

[3] DUDEN (2017): Resilienz. URL: http://www.duden.de/rechtschreibung/Resilienz [Stand: 08.06.2017]

[4] Edwards C (2009): Resilient Nation, London: Demos

[5] Flynn S (2011): A National Security Perpesctive on Resilience. In: Resilience: Interdisciplinary Perspectives on Science and Humanitarianism, 2, S. i-ii

[6] Goerger S, Madni A, Eslinger O (2014): Engineered Resilient Systems: A DoD Perspective. In: Procedia Computer Science, 28, S. 865–872

[7] heise.de (2017): WannaCry: Was wir bisher über die Ransomware-Attacke wissen . Url: https://www.heise.de/newsticker/meldung/WannaCry-Was-wir-bisher-ueber-die-Ransomware-Attacke-wissen-3713502.html [Stand: 08.06.2017]

[8] Holland J (2014): Complexity. A Very Short Introduction (Very Short Introductions). Oxford University Press, Oxford

[9] Holling C (1973): Resilience and Stability of Ecological Systems. In: Annual Review of Ecology and Systematics, 4, S. 1-23

[10] Hollnagel E, Fujita Y (2013): The Fukushima disaster – systemic failure as the lack of resilience. In: Nuclear Engineering and Technology, 45:1, S. 13-20

[11] Hollnagel E (2011): Prologue: The Scope of Resilience Engineering. In: Hollnagel E, Pariès J, Woods D, Wreathall J (Hrsg.): Resilience Engineering in Practice. A Guidebook, Farnham, Surrey: Ashgate, S. xxix-xxxix

[12] Kaufmann S, Blum S (2012): Governing (In)Security: The Rise of Resilience. In: Gander H, Perron W, Poscher R, Riescher G, Würtenberger T (Hrsg.): Resilienz in der offenen Gesellschaft. Symposium des Centre for Security and Society, Baden-Baden: Nomos, S. 235-257

[13] Linkov I, Kröger W, Renn O, Scharte B et al. (2014): Risking Resilience: Changing the Resilience Paradigm. Commentary to Nature Climate Change 4:6, S. 407–409

[14] Madni A, Jackson S (2009): Towards a Conceptual Framework for Resilience Engineering. In: IEEE Systems Journal, 32:2, S. 181-191

[15] Narzisi G, Mincer J, Smith S, Mishra B (2007): Resilience in the Face of Disaster: Accounting for Varying Disaster Magnitudes, Resource Topologies, and (Sub)Population Distributions in the PLAN C Emergency Planning Tool

[16] Nemeth C (2008): Resilience Engineering: The Birth of a Notion. In: Hollnagel E, Nemeth C, Dekker S (Hrsg.): Resilience Engineering Perspectives. Volume 1: Remaining Sensitive to the Possibility of Failure. Farnham, Surrey: Ashgate, S. 3-9

[17] Pescaroli G, Alexander D (2015): A definition of cascading disasters and cascading effects: Going beyond the „toppling dominos“ metaphor. In: GRF Davos Planet@Risk, 3:1, Special Issue on the 5th IDRC Davos 2014, S. 58-67

[18] Plodinec M (2009): Definitions of Resilience: An Analysis, Community and Regional Resilience Institute

[19] Scharte B, Hiller D, Leismann T, Thoma K (2014): Einleitung. In: Thoma K (Hrsg.): Resilien Tech. Resilience by Design: Strategie für die technologischen Zukunftsthemen (acatech STUDIE). München: Herbert Utz Verlag, S. 9-18

[20] Steen R, Aven T (2011): A risk perspective suitable for resilience engineering. In: Safety Science 49, S. 292-297

[21] telegraph.co.uk (2017): NHS cyber attack: Everything you need to know about ‚biggest ransomware‘ offensive in history. Url: http://www.telegraph.co.uk/news/2017/05/13/nhs-cyber-attack-everything-need-know-biggest-ransomware-offensive/ [Stand: 08.06.2017]

[22] The National Academies (2012): Disaster Resilience. A National Imperative, Washington, D.C.

[23] Thoma K, Scharte B, Hiller D, Leismann T (2016): Resilience Engineering as Part of Security Research: Definitions, Concepts and Science Approaches. In: European Journal for Security Research, 1:1, S. 3-19

[24] Wildavsky A (1988): Searching for Safety (Studies in social philosophy & policy, no. 10), Piscataway: Transaction Publishers

[25] Woods D (2003): Creating Foresight: How Resilience Engineering Can Transform NASA's Approach to Risky Decision Making. Testimony on The Future of NASA for Committee on Commerce Science and Transportation

Blockchain

18

Reliable Transactions

Prof. Dr. Wolfgang Prinz · Prof. Dr. Thomas Rose ·
Thomas Osterland · Clemens Putschli
Fraunhofer Institute for Applied Information Technology FIT

Summary

Blockchain technology has major relevance for the digitization of services and processes in many different areas of application beyond the financial industry and independent of cryptocurrencies in particular. Whilst for the Internet of Things the potential for automation associated with smart contracts is especially significant, for applications from the supply chain fields or for proofs of origin it is the irreversibility of the transactions conducted. This article describes the functioning of this new technology and the most important resulting qualities. The chapter provides a list of criteria for identifying digitization projects for which blockchain technology is suitable.

Because of the extent of blockchain technologies and their applications, developing the basic technologies requires a multidisciplinary approach, as does developing applications, carrying out studies of cost-effectiveness, and designing new governance models. The diverse competencies offered by the various Fraunhofer institutes, put the Fraunhofer-Gesellschaft in a position to make a significant contribution to the ongoing development and application of blockchain technology.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_18

18.1 Introduction

Trust and reliability are the critical key elements for the digitization of business processes, whether they take place between sales portals and customers or as inter-organizational processes between business partners working together within supply chains. While reputation management methods have attempted to use transactional analyses to support seller confidence in business-to-consumer and consumer-to-consumer relationships, today, in an Internet of Value, the question of confidence in transactions that depict different kinds of values immediately arises. Databases and process management have traditionally always pursued a centralized approach here, beginning with a nominated authority and central process synchronization. This centralization nevertheless entails a range of potential risks. These include for example performance bottlenecks, fault tolerance, authenticity, or internal and external attacks on integrity.

In the case of cryptocurrencies on the other hand, the central clearing function of banks is replaced by shared algorithms for ensuring correctness in the network. The central innovation of cryptocurrencies is thus guaranteeing the correctness of transactions within a network and also shared consensus finding between partners in the network. Consensus on the correctness of transactions and business processes is not managed centrally, but it is developed through shared consensus finding between the partners.

Since the publication of Satoshi Nakamoto's white paper in 2008 [3] and the creation of the first bitcoins in early 2009, both cryptocurrencies and blockchain technology have received ever more attention in the last two years. The reasons for this are the following features of the technology:

- Documents and investments can be uniformly encoded in a forgery-proof way and the transfer between senders and recipients can be stored as a transaction in the blockchain.
- The storage of the transactions is irreversible and transparent based on distributed consensus building and encryption.
- Transactions are verified within a peer-to-peer (P2P) network rather than by a central authority.
- Smart contracts offer the possibility of describing and executing complex transactions and assuring their boundary conditions. They enable both the automation of simple processes in the Internet of Things and new governance models by establishing alternative organizational forms.

Blockchain's potential applications thus extend far beyond cryptocurrencies. The technology is able to introduce a new generation of the Internet of Value or Trust after the Internet of Things. In what follows, the present chapter first describes the functioning of blockchain before illustrating the applications. An in-depth account of the technology and its application is provided in [8].

18.2 Functioning

The irreversible recording of transactions and the delegating of the sovereignty of a certifying authority to distributed consensus finding – both blockchain key features – are based on combining different techniques as shown in the simplified process illustration below.

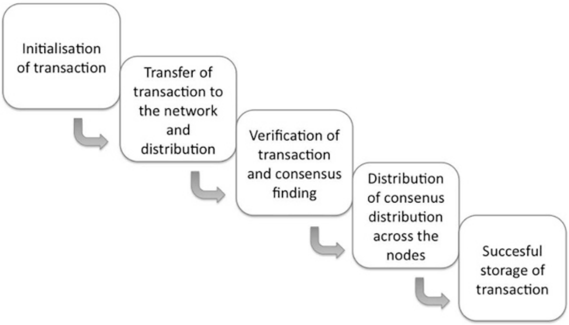


Fig. 18.1 Functioning of a blockchain (Fraunhofer FIT)

A core element of the technology is the encoding of transactions using hashing. Here, arbitrary character strings are converted into uniform encoding, where representation of different strings by the same code is precluded (collision resistance). An additional core element is consensus finding regarding the correctness of transactions. After transactions have been formally verified, partners within the network attempt to find a consensus regarding these transactions. Various procedures are used to find consensus. If a consensus is found for the transactions, they are distributed within the network and recorded in the global blockchain.

First, each transaction (e.g. a cryptocurrency transfer or document registration) is generated by a sender and digitally signed. This transaction is sent to the network and distributed to the participating nodes. The various nodes of the network verify the transaction's validity and attempt to find a consensus for the entire block, which includes this transaction. Next, the "mutually accepted" consensus is broadcasted across the network and accepted by the nodes, thus extending the blockchain. Transactions reside in blocks and all types of transactions are converted into a standardized format using hash functions. To do this, all the individual transactions are encoded into hash values and then compressed hierarchically. This hierarchical compression is known as a hash tree or Merkle tree, which allows a block of transactions to be represented unambiguously. This encoding is secure against attempts at manipulation since changing even one transaction would change the hash value of the block and the hash tree would thus no longer be consistent.

Blocks are linked to preexisting blocks via concatenation to produce a (block) chain. For a block to be accepted as a new element into the existing concatenation, a process of consensus building as described in section 18.3 must take place, resulting in a correct and irreversible concatenation of blocks to form a blockchain. To ensure persistence, these chains are replicated in all nodes of the network; that is, all nodes have the same basic knowledge.

Blockchains can thus be described in a simplified form as distributed databases that are organized by the participants in the network. In contrast to centralized approaches, blockchains are far less prone to error. Nevertheless, these systems also entail various challenges. At present, critical discussions focus in particular on the high data redundancy, since holding multiple copies of the same data within the network requires a very large amount of storage space.

18.3 Methods of consensus building

Consensus building is an essential and fundamental pillar of the blockchain concept. It validates transactions in such a way that agreement can be reached on which transactions are to be recognized as valid, where the saved declaration acknowledged by everyone is immutable in future. This method of distributed systems is also known as a solution to the Byzantine Generals Problem. This is concerned with identifying whether messages remain authentic and unaltered as they travel between different recipients. The processes utilized here are based on concepts that have long been the subject of research in the context of distributed networks [2] and distributed systems [6].

The currently best-known example of a blockchain implementation process used is the bitcoin blockchain proof-of-work. Interestingly, the actual proof-of-work concept was already suggested as far back as 1993 for stemming the tide of junk email [5]. It is based on an asymmetric approach where a service user (in this case, the sender of an email) has to complete work that can be verified relatively easily by a service provider (in this case, the email network provider). That is, only those who perform work on behalf of the community are permitted to also use the community's resources. In the blockchain context, the users are the miners who perform the difficult task of computing the proof-of-work, and the providers are all of the nodes who carry out relatively straightforward checks to see whether the successful miner has computed the proof-of-work properly. In the bitcoin blockchain, the proof-of-work algorithm is based on the hashcash process presented by Adam Back [1]. The goal of the algorithm is to find a number (nonce, number used only once) that, when combined with the new block to be attached to the existing blockchain, produces a hash value that fulfills a specific condition. One example condition is that the value to be identified must consist of a specific number of leading zeros. This number can only be identified by means of extensive trial and error as hash functions are one-way functions.

In this particular proof-of-work process, the computing power of the nodes is a significant factor in who solves the problem and identifies a suitable nonce value. Since miners are rewarded with new bitcoins for identifying the nonce, a competition is created among miners where they invest in ever-increasing computing power. This would decrease the time required to find a valid nonce. However, this is inconsistent with the bitcoin network's rule that a new block should only be generated approximately every 10 minutes, which is related to the fact that successful miners are rewarded for "freshly minted" bitcoins. If the intervals between new blocks being created were shortened, then the total amount of money in circulation would grow too quickly. For this reason, the difficulty level of the puzzle is increased whenever the period of time is shortened by the addition of new processing capacity. For miners operating the computing nodes, this means increased work with decreased prospects of success.

Since the work involved consists primarily of the energy used, alongside the investment in computing power, the proof-of-work approach does not make sense for all blockchain applications. This is particularly the case for applications where this kind of competition is unnecessary. Therefore, alternative proof-of-work processes were developed that are either memory or network-based. In the case of memory-based approaches, the puzzle is solved not by computing power but by a corresponding number of memory accesses [1], whereas in the network-based approach, by contrast, it is solved only by communication with other network nodes (e.g. to gather information from them that is required to solve the puzzle) [9]. The proof-of-work process makes sense when access to the blockchain network is public and not subject to any access restrictions. An alternative process, primarily relevant for private blockchains (see Ch. 18.4), where the nodes involved are known and subject to access restrictions, is the proof-of-stake process. Here, nodes that are able to validate a new block are selected according to their shares in the cryptocurrency [5] or via a random process [7]. The selection of the most suitable process is dependent on the specific use case and the blockchain solution used. An additional important aspect is scalability for transaction volumes, especially in the case of applications in the Internet of Things. Current approaches are not able to compete with databases with respect to their frequency of transactions. This aspect, as well as the fact that in a blockchain all of the data stored is replicated in each node, mean that blockchain solutions initially cannot be used for data storage in the same way that databases can. They take on special tasks in combination with databases when the focus is on managing information reliably, with common agreement on the transactions to be recognized as valid. In addition, databases store the payload, while a fingerprint of the data is filed in the relevant blockchain to guarantee integrity.

18.4 Implementations and classification

The following section provides a classification of different blockchain system implementations and differentiates between various conceptional models.

The key classification for blockchains is the degree of decentralization of the entire network. This degree is determined by means of various properties: it starts from a traditional central database and ends with a completely distributed blockchain. For this reason, each blockchain system can also be viewed as a distributed database.

Blockchains may, like databases, be available privately or publicly. The primary distinction however is made by who may use the system, that is, which user is permitted to add new transactions to the blockchain. If the user requires the permission of an organization or a consortium, it is a private blockchain. If, however, every user is allowed to write new information into the blockchain then it is public.

For a public blockchain, we also need to distinguish who is permitted to summarize the newly added transactions into blocks and validate them. In a *permissionless* system, every user can add new blocks and validate them. Normally, economic incentives are given for this so that users behave appropriately. The user may for example receive specified transaction costs from the transactions contained in the blocks.

In a blockchain system requiring authorization (*permissioned*), only particular users are permitted to add and validate new blocks. These users are identified by an organization or consortium. The shared trust process is thus only distributed to those users who have been authorized by the consortium, however, and not to all partici-

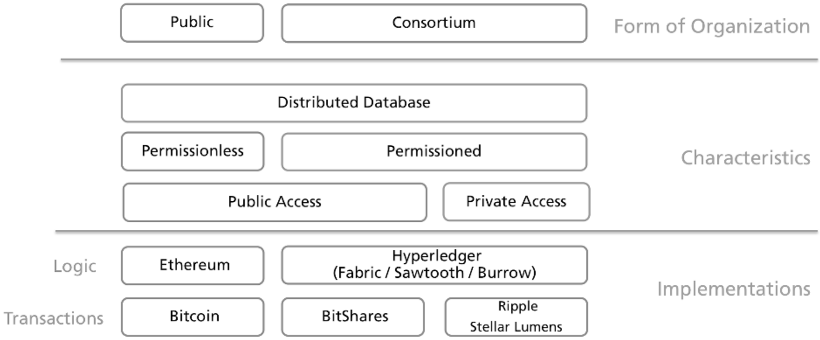


Fig. 18.2 Classification of blockchains (following [4])

pating users. For validation, authorized individuals normally carry out a simplified consensus process (e.g. proof of stake) that may be far more efficient.

Blockchain implementations may additionally be differentiated by the extent to which they are oriented as a platform around solving logical problems, or whether they are rather designed for "traditional" cryptocurrencies. On some blockchain implementations such as Ethereum or Hyperledger Burrow, Turing complete smart contracts can be executed, for example, that is, smart contracts may be complex programs instead of simply conditional transactions. Fig. 18.2 contrasts the various classification properties and lists a number of important blockchain implementations for comparison.

18.5 Applications

The terms "smart contract" or "chaincode" refer to programs that are executed within the blockchain network. Once smart contracts are saved and instantiated in the blockchain, they are immutable, and the execution of the processes defined in the program code is independent of external entities. A smart contract can become active via an external event or a user interaction.

The immutability of instantiated smart contracts and the additional option of modeling complex processes permit the reliable handling of transactions between various entities. The cryptographically secured execution of smart contracts in the blockchain means they are not only beneficial for carrying out a defined process but they also simultaneously document the process itself.

One example of a concrete application is smart grids, which represent a significant change from current power supply systems. They transform a centralized organizational structure, which is shaped by a small number of large power generators such as coal or nuclear power plants into a network with many inhomogeneous small generators such as solar arrays and wind turbines. In this kind of network, a gardening enthusiast wanting to mow their lawn could buy the required electricity directly from their neighbor's solar array. A smart contract acting for the solar array in the energy market would be the interface that contacts the gardening enthusiast's local power supply – also represented on the energy market by a smart contract. In the process, smart contracts are automatically able, within regulatory limits, to negotiate an electricity price and calculate the energy purchased. To do this, the solar array's smart contract verifies that the correct sum is received for every kilowatt hour purchased, while the gardening enthusiast's smart contract checks that the correct amount of power is received for the sum paid.

Where blockchain's characteristic qualities are carried over to computer programs, smart contracts become an interesting alternative for application fields where critical intermediaries can be replaced by programs that are clearly defined and operate transparently. These qualities allow not only the automated, reliable initiation of transactions within the blockchain (due to their distributive, independent execution), but also serve to maintain consistency between different entities connected by the blockchain.

Blockchain solutions have significant potential within specific boundary conditions if one or more of the following criteria are fulfilled.

1. Intermediaries: in the use case in question, intermediaries in the process can or should be avoided. Companies should thus examine their processes and business models to see if they could either fulfill the role of intermediary themselves or otherwise optimize processes where they are reliant on an intermediary. Using a blockchain makes sense when
    a. the intermediary creates costs for the process steps that could just as well be provided by blockchain functions
    b. the intermediary delays a process and a blockchain application could speed it up
    c. political reasons favor changing from central, intermediary-managed processes to decentralized ones.
2. Data and process integrity: retrospective immutability and precisely specified implementation of the transaction are required for this use case.
3. Decentralized network: utilizing a network of validating or passive income nodes that carry out processes autonomously makes sense and/or is possible. This is relevant for all processes that involve flexible, new, and fleeting cooper-

ation partners without a stable and secure basis for transactions and trust. In these cases, a blockchain can guarantee networked integrity.

4. Transferring value and protecting rights: blockchains facilitate the transfer of value and rights. Thus, all processes are relevant where original copies, proofs of origin, or rights need to be conveyed or transferred.

In addition to these criteria, it is important that the focus should not lie on the use of a cryptocurrency itself and that no processes that are subject to strict regulation should be selected for an initial assessment of the technology and the development of demonstrators.

Sources and literature

[1] Adam Back 2002. Hashcash – A Denial of Service Counter-Measure.
[2] Baran, P. 1964. On Distributed Communications Networks. IEEE Transactions on Communications Systems. 12, 1 (Mar. 1964), 1–9.
[3] Bitcoin: A Peer-to-Peer Electronic Cash System: 2008. https://bitcoin.org/bitcoin.pdf. Accessed: 2017-03-16.
[4] Distributed Ledger Technology: beyond block chain: 2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/492972/gs-16-1-distributed-ledger-technology.pdf. Accessed: 2017-06-27.
[5] King, S. and Nadal, S. 2012. Ppcoin: Peer-to-peer crypto-currency with proof-of-stake. self-published paper, August. 19, (2012).
[6] Lamport, L. et al. 1982. The Byzantine Generals Problem. ACM Trans. Program. Lang. Syst. 4, 3 (Jul. 1982), 382–401.
[7] Whitepaper:Nxt: https://nxtwiki.org/wiki/Whitepaper:Nxt.
[8] Wolfgang Prinz and Alexander Schulte (Eds) 2017. Blockchain: Technologien, Forschungsfragen und Anwendungen – Positionspapier der Fraunhofer Gesellschaft. to appear
[9] Znati, T. and Abliz, M. 2009. A Guided Tour Puzzle for Denial of Service Prevention. Computer Security Applications Conference, Annual (Los Alamitos, CA, USA, 2009), 279–288.

E-Health

19

Digital Transformation and its Potential for Healthcare

Prof. D.Eng. Horst Hahn · Andreas Schreiber
Fraunhofer Institute for Medical Image Computing MEVIS

Summary

While the digital transformation is well underway in numerous areas of society, medicine still faces immense challenges. Nevertheless, the potential resulting from the interaction of modern biotechnology and information technology is huge. Initial signs of the transformation can be seen in numerous places – a transformation that will further be accelerated by the integration of previously separate medical data silos and the focused use of new technologies. In this chapter, we describe the current state of integrated diagnostics and the mechanisms of action behind the emerging field of digital healthcare. One of the areas of focus is the recent revolution caused by artificial intelligence. At the same time, we have seen the emancipation of patients who now have access to an enormous breadth of medical knowledge via social networks, Internet search engines, and healthcare guides and apps. Against this backdrop, we will discuss the change in the doctor-patient relationship as well as the changing roles of doctors and computers, and the resulting business models.

19.1 Introduction

The digital transformation that is currently the topic of discussion in every segment of the market and technology has barely taken place in the field of healthcare. Simultaneously, we are observing a rapid increase in complexity right across the medical specialties that for years now has been stretching the limits of what is fea-

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_19

sible for those working in them. The following commentary is based on the hypothesis that digital transformation is the key to raising healthcare to the next level in terms of success rates, security, and cost-efficiency, while simultaneously realizing the potential of modern medicine. The distribution of responsibilities within the medical disciplines is just as much under scrutiny as remuneration mechanisms for enhanced healthcare services, applicable quality standards, medical training curricula, and, last but not least, the role of empowered patients.

One of the preconditions for this transformation, the digitization of all relevant data, is already at an advanced stage: most patient information collected in this country is already available in digital form. An exception is clinical pathology where tissue samples generally continue to be assessed manually under optical microscopes. At the other end of the spectrum, process automation in laboratory medicine has long been part of the status quo. What remains neglected, however, is the connection of the individual sectors as well as the structured use of integrated information.

The digital transformation is supported by the interplay of several apparently independent technologies that have become remarkably powerful in recent years (cf. Fig. 19.1). On the one hand these are developments outside of medical technol-

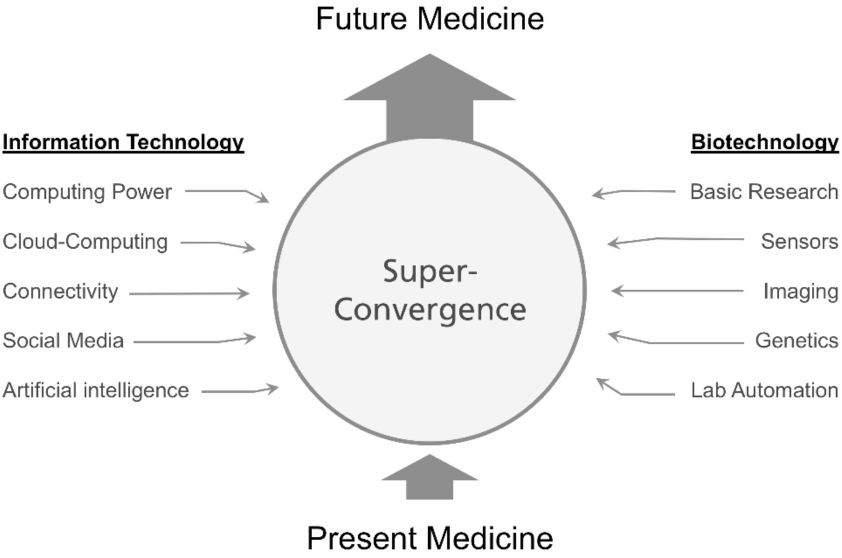


Fig. 19.1 Technological super convergence as a precondition for digital transformation in healthcare, according to [26] (Fraunhofer MEVIS)

ogy: the sheer computing power and storage capacity of modern computers of all sizes with constantly increasing network bandwidth and cloud computing, as well as far-reaching connectivity via the Internet, mobile devices, and not least social media and artificial intelligence. On the other hand these are achievements in biotechnology, laboratory automation, microsensors and imaging, as well as, very often, the findings of basic medical research. In a matter of just ten years, it has been possible to reduce the cost of sequencing an entire genome by a factor of six to just a few hundred Euros. Eric Topol, in his book *The Creative Destruction of Medicine* [26], describes these simultaneous developments as a “super convergence” from which the new healthcare arises.

In what follows, we describe the mechanisms of action of this digital healthcare at its various stages from prevention and early diagnosis through to clinical treatment. Thus it is possible, even now, to see numerous cases of integrated diagnostics with new business models. We then discuss the revolution in artificial intelligence being observed across every field (including healthcare), and the changing roles of doctors and patients and of the different medical specialties. From a higher-level perspective we also discuss the health-economic potential of digital medicine and the changing industry landscape, where the battle for sovereignty over data integration and data access is increasingly emerging. The last section provides a brief outlook, touching on additional and related subjects that are treated elsewhere due to the brevity of this present text.

19.2 Integrated diagnostics and therapy

19.2.1 Digitization latecomers

In the first decades of the 21st century, digital transformation is already in full swing and is exercising far-reaching influence on numerous areas of society. Smartphones and ubiquitous mobile access to the Internet in particular have changed the way we communicate, work, access information, teach and learn, and – very importantly – how, what, and where we consume. In the process, the gradual digitization of everyday life has increased the efficiency of many processes and democratized access to information and knowledge with the aid of participative platforms. Most of the time, this has strengthened the role of the consumer thanks to increased transparency. At the same time, however, the “data power” of large corporations is growing, along with the risk that data will be manipulated or utilized to the disadvantage of the user.

Digitization presents society and its stakeholders with key challenges, since it accelerates institutional change and demands a high degree of flexibility from everyone involved. An example with numerous similarities to medicine would be the media landscape, where digitization gave rise to new distribution channels and empowered consumers, forcing a new market order into being that redefined the role of established media producers. The resulting new diversity of media and faster publication cycles, however, also increase the chances of manipulation and may make it more difficult to receive high-quality information.

The disruption caused by digitization within manufacturing is no less far reaching and extends to the complete redesigning of industrial value chains. Frequently referred to as "Industry 4.0", the digital transformation of production aims to achieve a high degree of process automation together with the autonomous sharing of data between machines in order to guarantee seamless processes across manufacturing and supply chains.

While the products and services we consume in the digital age are increasingly tailored, healthcare is still dominated by the one-size-fits-all principle, closely allied to the principles of evidence-based medicine. Already today, the majority of medical data is captured digitally and, alongside record-taking and patient management, also selectively used to support clinical decision-making such as diagnosis and therapy selection.

Nevertheless, the true potential of broad computer-assisted medicine is going to waste since data is at times documented and saved incompletely and in an insufficiently standardized format, and a lack of suitable interfaces as well as outdated legal contexts prevent centralized storage in the form of electronic patient records. The latter is a necessary precondition both for future support in personalized patient decision-making as well as for the preliminary assessment of large medically curated databases for training self-learning algorithms [11].

19.2.2 Innovative sensors and intelligent software assistants

This is where the medicine of the future will come to our aid. Alongside advances in genetics and molecular biology as well as in diagnostic and interventional medical technology, digitization is a pivotal foundation for future "personalized medicine", also known as "precision medicine". Here, intelligent software solutions function as integrators of all the relevant patient information and fields of care and are the keys to integrated predictive diagnostics and therapy planning (cf. Fig. 19.2).

In future, personalized medicine will be based on individual patient risk assessment in view of family history, genetic predisposition as well as socioeconomic and

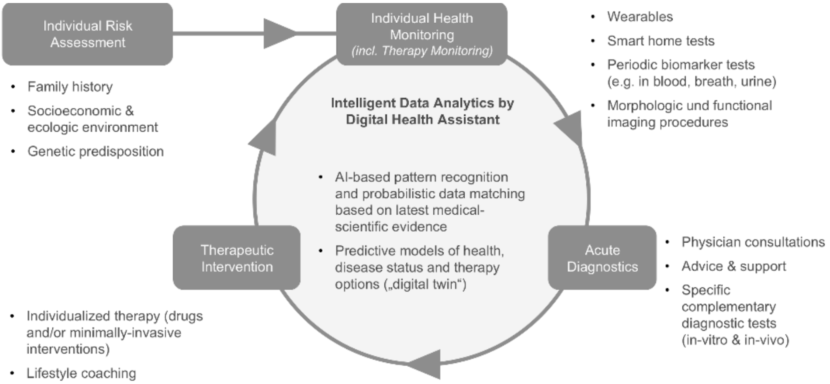


Fig. 19.2 Information loop of personalized medicine (Fraunhofer MEVIS)

ecological environmental parameters, providing a multi-factor risk score, which is adjusted to changing environmental factors. In particular, this risk assessment will benefit from advances and the improved cost-efficiency of genome and molecular diagnostics. More than 200,000 gene variants are known to be associated with the development of individual illnesses, from a human genome of around 20,000 genes [22]. While just a few of these mutations feature a specific, high disease risk, it is expected that research will continue to uncover additional clinically relevant relationships between different gene variants, and between variants and environmental influences [19].

19.2.3 Population research

Against this backdrop the motivation also arises to incorporate the systematic and long-term recording of environmental factors such as fine particulate air pollution or the geographic distribution of viral infections into personal healthcare provision and therapy planning. Without computer assistance, these complex fields of data and knowledge cannot be exploited by doctors, since intelligent analysis systems are required that merge regular knowledge with statistical models and are thus able to generate individualized patient risk assessments.

Several projects were commenced in recent years that seek to gather comprehensive health data (including imaging) from right across the population and make it accessible for disease risk investigation. The goal in each case is to analyze the health profiles of subjects over the long term in order to identify and interpret early

Fig. 19.3 The GNC Incidental Findings Viewer is accessible via the Internet and contains a structured database of findings as well as automatic image quality analysis. (Fraunhofer MEVIS)

indications and reasons for diverse ailments. The most recent of these is the German National Cohort (GNC), with 200,000 planned participants, of whom around one in seven even undergoes an extensive whole-body MRI (see Fig. 19.3). Other similar initiatives are the UK Biobank with 500,000 participants and a US study conducted as part of the Precision Medicine Initiative. Also of particular interest here is Project Baseline, a joint endeavor of Stanford and Duke universities which, alongside annual examinations, is also collecting data via passive sleep sensors and smartwatches.

19.2.4 Multi-parameter health monitoring

As a complement to present risk assessment approaches, modern sensors offer the opportunity to continuously or periodically measure key vital parameters, for example, and evaluate them for early diagnostic detection. Wearables are the first class of devices we should mention here, and they may be capable of capturing and evaluating pulse rates and temperature or even oxygen saturation and glucose levels in the blood. The digital assistant from Fraunhofer's *&gesund* spin-off, for example, uses the sensors of common smartwatches to define personal normal levels and detect any deviations. This multi-parameter monitoring is already being tested for the early detection of heart disease and to assist in the treatment of lung diseases and mood disorders.

Regular measurement and evaluation of disease-specific biomarkers in the blood and other bodily fluids are also expected to provide early indications of the development of conditions and of response to therapy. For example, liquid biopsies have recently led to successful tumor identification via the use of corresponding antibodies to detect tumor DNA circulating freely in the blood. Proteome and metabolome are also becoming of increasing interest to medicine since they are credited with significance for overall health and the development of various diseases.

In future, these and other analyses will take place unnoticed alongside traditional blood tests. Researchers at Stanford University's Gambhir Lab, for example, are working on lavatory-based detectors that automatically test stool samples for pathogens and indications of diseases [8]. And in the near future, we will be able to carry out a range of in vitro tests in our emerging smart homes, even though still today these tests require costly laboratory infrastructure. It also remains to be seen how long it will take for the analysis of behaviors, speech, gestures, and facial expressions to contribute to the early detection of affective and other psychotic conditions. Duke University's *Autism & Beyond*[1] research group, for example, promises to detect signs of autism in early childhood based on automated video diagnostics.

Medical imaging will continue to play a significant role in personalized early detection. Although large-scale screening programs, e.g. for the early detection of breast cancer, are being vigorously debated today, imaging remains a procedure with high specificity. In future, we will need to further stratify the patient populations benefiting from imaging and in each case employ the best possible imaging workup to significantly minimize false positive results and side effects. Magnetic resonance imaging and ultrasound in particular appear to be increasingly used in screening.

Whereas MRIs will become even faster and more cost-efficient in the coming years, sonography is benefitting from hardware and software advances that allow for high-resolution, spatially differentiated and quantifiable diagnosis. The fact that neither modality emits ionizing radiation represents a key safety advantage over computed tomography and traditional projection radiography, despite successes in radiation dosage reduction.

The personalized health monitoring expects that individuals increasingly assume personal responsibility for their health and develop an awareness of the key lifestyle factors. App-based guides on mobile devices contribute towards this by developing tailored recommendations for nutrition and sporting activities. Apps such as *Cara* for gastrointestinal complaints, *mySugr* for diabetes, and *M-sense* for migraines are all designed to assist personal health management.

---

[1] Autism & Beyond: https://autismandbeyond.researchkit.duke.edu/

19.2.5 Digitization as a catalyst for integrated diagnosis

In case of acute conditions, we expect an even closer linking of monitoring and care with the traditional diagnostic processes, the effectiveness of which will also continue to increase in the coming years. Results from laboratory diagnostics and imaging are in most cases already available in digital form today, and there is a trend towards connecting those data silos, often currently still separated, via appropriate interfaces. In ultrasound diagnostics, too, where analysis is still primarily carried out live at the examination venue, an increasingly standardized image acquisition process and centralized storage will form the basis in future. In the USA, storage and appraisal often take place separately in the division of labor of sonographers and radiologists.

Clinical pathology, where tissue samples are examined and analyzed under the microscope, still operates the conventional way. Alongside traditional histological assessments, modern immunohistochemistry and molecular pathology processes yield more specific but also complex patterns. In future, slide scanning and consistent digital characterization of tissue samples will play a significant role in this rich information being processed in everyday clinical practice.

Virtual multiple staining, for example, allows several specific serial sections to be combined so the overall tissue information contained can subsequently be analyzed automatically or semi-automatically (see Fig. 19.4). Examination under the microscope does not even permit the visual precision correlation of two stains

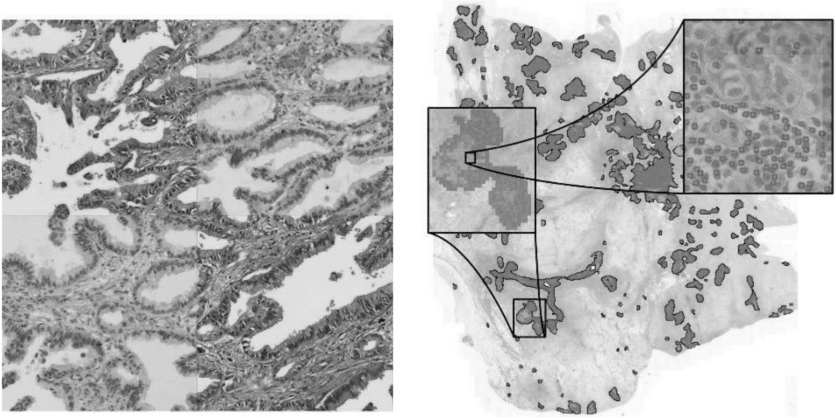


Fig. 19.4 Left: virtual multiple staining using high-accuracy serial section image registration. Right: automatic analysis of lymphocytes in breast tissue images (Fraunhofer MEVIS)

Fig. 19.5 High accuracy simulation for tumor therapy: for radiofrequency ablation, the expected temperature surrounding the probe is calculated (© Fraunhofer MEVIS)

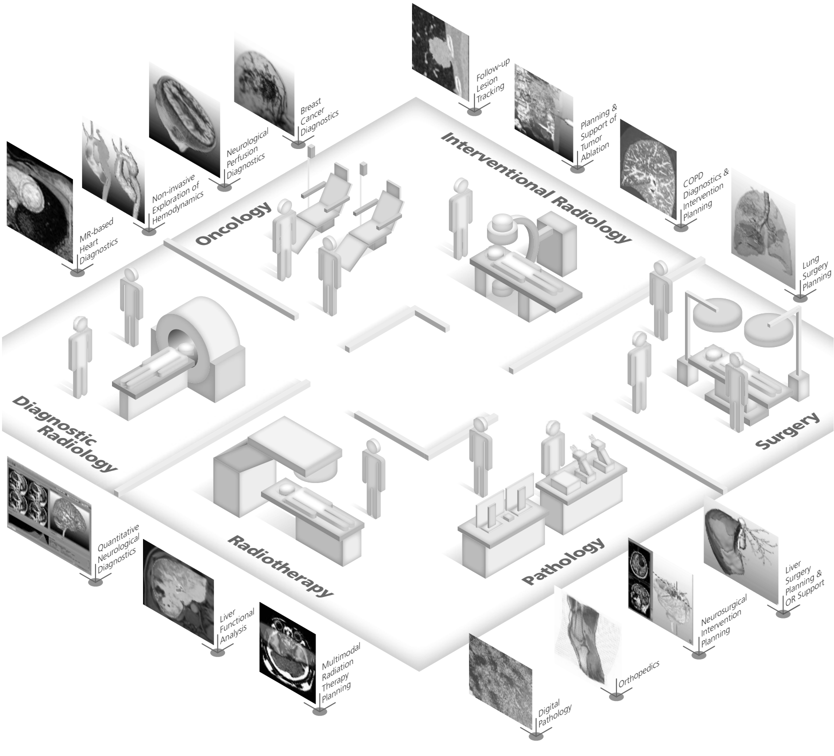


Fig 19.6 Imaging plays a central role in nearly every clinical discipline and contributes significantly to high-precision, integrated care. (Fraunhofer MEVIS)

(Fig. 19.4 ii). With subsequent generations of technology, the persistent deficits in image quality and time required are expected to give way to an immense increase in objectivity and productivity, particularly in the case of recording quantitative parameters. We expect, as with the digitization of radiology, that pathology will entirely adopt digital processes within just a few device generations and experience a standardization of methods and reporting.

Nevertheless, the true catalyst of diagnosis is information technology that for the first time allows us to consider diagnosis based on integrated data right across disciplines. By "integrated diagnostics" we mean using software to bring together all of the diagnostically-relevant information – from the laboratory, radiology, pathology, or the individual health and case files – to allow statistical comparison of the biomarker profile and, finally, intelligent differential diagnosis and decision-making support. The "digital twin" that results is not to be understood literally but is based on precisely this information integration and permits the predictive modeling of potential courses of disease and the probabilistic prioritization of therapy options.

Imaging data here plays an important role in phenotyping, planning of interventions, and in detailed therapeutic monitoring (cf. Fig. 19.7). Innovative approaches also permit the direct utilization of 3D planning data in the operating room. The increasing dissemination of robotics in surgery and interventional radiology will also further boost the significance of intraoperative imaging, because mechatronic assistance systems achieve their full potential in particular in combination with precise real-time navigation and corresponding simulation models (cf. Fig. 19.5 and 19.6)

19.3 AI, our hard-working "colleague"

19.3.1 Deep learning breaks records

The transformation towards highly efficient digital medicine will only occur when it is possible to analyze and interpret the exponentially growing data volumes efficiently. Artificial intelligence (AI) and machine learning methods – which have been revolutionized in just these last few years – thus play a crucial role. In educated specialist circles and beyond, deep learning is on everyone's lips. Dismissed as a hype by some, there is nevertheless extensive agreement that a level of maturity has been achieved that allows the most complex practical problems to be effectively solved.

AlphaGo's victory in early 2016 [20] over the world's best *Go* players is just one example of how the tide has turned in a short space of time. Just a few years ago, the common belief was that computers would require several decades before they could play *Go* at the level of a grand master – if ever, in view of the enormous diversity of variations. The theoretical total number of possible board positions is a figure with more than 170 digits. Even if we deduct the unrealistic positions (a large proportion) from this, we are still left with a figure that makes the approximate number of atoms in the universe (around 10^{81}) appear vanishingly small.

Public AI consciousness was boosted in 2011 when IBM's Watson computer beat the best players ever at the time, Ken Jennings and Brad Rutter [10] by an immense distance, in the quiz show *Jeopardy!* Technically, Watson is just a huge database with a search engine that not only recognizes logical but also *meaningful* connections in the data, and a so-called natural language processor that is able to understand our spoken or written language. At the time, the Watson database comprised around 200 million pages of text and tables – the system built on it could thus answer the most commonly posed questions reliably even without an Internet connection. The understanding of language alone has undergone a revolution due to deep learning, and today Alexa, Siri, Baidu et al. along with the latest translation machines understand human language nearly as well as humans [17].

19.3.2 Pattern recognition as a powerful tool in medicine

More relevant, perhaps, for medical applications are the successes of so-called "convolutional neural networks" (CNNs), a special variant of deep neural networks where the individual weight factors of multilayered linear convolution operations are learnt from the training data. As a result, the network adapts itself to those visual features that have the most significance for a given problem. In this context, the network's "depth" refers to the number of layers.

The breakthrough took place in 2012 in the context of the annual ImageNet competition[2], based on the eponymous image database of more than 14 million photographs corresponding to over 21,000 terms. The goal of the competition is to identify the most suitable term for a random image selection. Up until 2011, the accuracy of automated computer systems had largely plateaued at an error rate of over 25%, far inferior to the approx. 5% of the best human experts. Once Alex Krizhevsky had used CNNs in his AlexNet to achieve a huge leap in accuracy in

---

[2] ImageNet Database, Stanford Vision Lab: http://www.image-net.org/

2012, all of the top ten places were taken over by deep learning/CNN approaches already by 2013. The next leap took place in 2015 when a team at Microsoft increased the network depth from 22 to 152 layers, thus breaking the barrier of human competence for the first time with an accuracy level of 3.7%.

The ideas behind deep learning as an extension of artificial neural networks (ANNs) have been around for several decades now, and their successful application in medicine was described more than 20 years ago [18, 28]. However, ANN research grew quiet in the late 1990s, unable to keep pace with the expectations awakened, in the face of simpler and more readily understood classification processes. The breakthrough of the recent years took place due to the exponential growth in computing power and in particular due to the use of graphics processors for solving the kinds of complex numerical optimization problems arising from deep learning and CNNs. Now, the first places in comparative medical image analysis competitions are thus almost completely occupied by the various CNN variants.[3]

In what could almost be described as a kind of gold rush, the newly discovered methods are being targeted at practically every problem within medical data analysis, and corresponding startups are springing up all over the place [4]. If the niche field had not caught their attention before, it was probably the $1 billion takeover by IBM of Merge Healthcare in summer 2015 that finally moved AI right up the CEO agenda of major medical technology groups worldwide.

19.3.3 Radiomics: a potential forerunner

At this point in time, integrated diagnosis shows most promise in *radiomics* applications. It combines phenotyping based on a large number of image-based quantitative parameters with the results of genome sequencing, laboratory findings, and (in future) even multisensory data from wearables and other sources. The goal is not only the simple detection of patterns that humans, too, would be able to recognize in the data, but the prediction of clinically relevant parameters such as drug therapy response [14]. Overall, radiomics is the vehicle for providing machine learning and integrated diagnostics with concrete demonstrable significance for solving complex clinical problems.

Thus, in the case of cancer therapy, for example, the combination of radiologically identified tumor heterogeneity (cf. Fig. 19.7 and 19.8) and specific laboratory parameters or results from molecular diagnostics could be decisive in discontin-

---

[3] COMIC – Consortium for Open Medical Image Computing, Grand Challenges in Biomedical Image Analysis: https://grand-challenge.org/

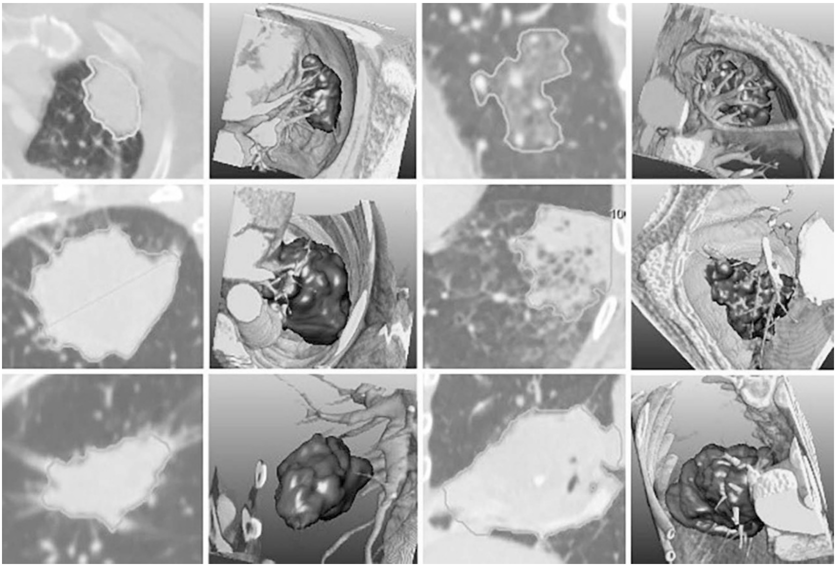


Fig. 19.7 Image-based phenotyping using computed tomography scans of six different lung tumors (Fraunhofer MEVIS, all rights reserved. CT data: S. Schönberg & T. Henzler, Mannheim)

uing a therapy or starting a new one [2]. Ultimately, software needs to be able to prioritize the huge volume of patient health information in specific individual cases and make it useable for clinical decision-making. Patient selection and precise choice of therapy are also especially essential for highly specific immune therapy, a form of treatment which has already demonstrated impressive results in recent years [3].

The AI techniques described above will be a great help in comparing individual data with population-specific databases. And more and more examples of diagnostic or prognostic tasks are coming to light where computers and people are meanwhile at least on a par with one another. One of these is the visual assessment of skin cancer based on high-resolution photographs. At the beginning of 2017, for example, researchers at Stanford University were able to demonstrate how a CNN trained on 129,450 images achieved outcomes similar to 21 certified dermatologists when classifying the most frequently occurring varieties of skin cancer [9].

19.3.4 Intuition and trust put to the test

Whether playing *Go* or supporting the medical decision-making process, computers develop a kind of "intuition" during the learning process that we previously did not expect from such logical and rigidly wired computer systems. They make predictions even if the algorithm cannot compute all possible combinations.

That is, when a given pattern only "feels" like it would lead to victory or belong to a specific medical category. Our inability to fully explain the responses of neural networks has already been widely discussed as a problem and is seen as a hurdle to the introduction into medical practice [13].

Should we trust a computer, even if we do not obtain a definitive explanation for its response? This is a question that leads to a deeper engagement with the self-learning nature of these kinds of AI systems. "Self-learning" means that these deep neural networks are able to generate their understanding based purely upon sample data and do not require the provision of additional explicit rules. Whereas simple features are extracted at the first levels of a network, at higher levels the patterns learnt are often highly specific and almost impossible to describe completely; it is these which help systems achieve their high level of accuracy.

In the meantime, researchers have managed to coax a visualization of the most relevant patterns in each case from trained networks, something which could help users to develop *trust* in the system as well as to discover malfunctions or false alarms. On closer examination, however, we see that a key part of the explanation behind the computer results remains hidden in a similar way to hu-

Fig. 19.8 Deep learning segments the liver and liver lesions. Left: downstream classifiers sort out false positive findings (dotted line), CT data: LiTS Challenge. Right: deep learning distinguishes tumors (striped) from cysts (Fraunhofer MEVIS, all rights reserved. CT data: R. Brüning, Hamburg).

man gut feeling, also often likely learnt through experience but equally hard to put into words.

And herein lies a key difference between people and computers, since trained CNNs can be comprehensively described and statistically validated, a feature that plays a significant role in authorization for use as a medical product. For the wider adoption of AI-based assistants it will also be essential that their inner workings are made comprehensible as much as possible and that remaining errors are analyzed thoroughly.

Time and again, we are likely to see that people and machines have greatly differing sources of error as well as strengths, in keeping with their very different learning and behavioral mechanisms. The correct design of user interfaces will thus also be key to the successful optimization of human/computer teams.

19.4 Changing distribution of roles

19.4.1 Integrated diagnostic teams

With the spread of specialist digital systems, large databases, and integrated diagnostic solutions, the distribution of roles between the medical specialties will change as well. We can assume that the extensive separation of working processes in radiology, pathology, and other diagnostic disciplines will be combined into an equally integrated cross-disciplinary working process.

The decisions made by tumor boards organized in numerous locations today, based on the findings of individual specialties, will in future be made right from the start by interdisciplinary diagnostic teams with extensive computer assistance, in a highly efficient and personalized manner. In the ideal world, all relevant information would flow together to these teams, which would represent a kind of control center for the medical decision-making process. As a result, seamless cooperation with the relevant neurologists, surgeons, cardiologists, oncologists, radiotherapists, gynecologists, urologists, etc. will take on an even greater significance than at present. Connectivity will thus not only become a key competitive factor with respect to data, but also with respect to the players involved.

19.4.2 The empowered patient

The transformation towards digitized medicine outlined here is also accompanied by changing doctor-patient interaction based on telemedicine or even intelligent medical chatbots. In future, we expect that a majority of medical consultations will take place virtually. In the USA, more than 5 million medical consultations are expected to take place via videoconferencing by 2020 [24]. This trend is triggered by both healthcare providers and payors, with a view on increased cost efficiency as well as patient comfort. Alongside general medical consultations, emergency medicine in particular will also benefit from telemedicine in order to efficiently access specialist knowledge in difficult and unusual cases.

The sheer availability of detailed specialist information is changing the structure of the doctor-patient relationship even more importantly than telemedicine. This can already be seen from the way doctors are regularly confronted with extensive partial knowledge in their practices, knowledge that patients have generally drawn from Wikipedia, from the various online healthcare advice sites, or from Dr. Google, their search engine for symptoms. Simply shutting out this reality with an exclamation such as, "If you're such an expert already, you certainly don't need my help!" would be inappropriate and a completely missed opportunity.

But this is just the beginning. Fred Trotter [27] described e-patients as the "hackers of the healthcare world". The "e" here stands for various adjectives, including "educated", "engaged", and "electronic", but, above all, "empowered". E-patients are the key actors behind "participative medicine"[4], a concept that has been propagated for around ten years. There are now a whole range of patient portals such as PatientsLikeMe and ACOR[5], where those afflicted can connect with one another and exchange rich illness-related information with each other and with doctors. For rare conditions in particular, the Internet is increasingly a better resource than general physicians.

The undeniable trend towards self-diagnosis and more empowered patients also requires discussion with respect to its potential dangers. Weak points arise due to authentication and data integrity issues related to shared information entry as well as due to overdocumentation with the potential for overdiagnosis/misdiagnosis and, finally, undesirable clinical results. Corresponding training for professionals, improved infrastructure, and reasonable legal frameworks can help to avoid these consequences. The various actors will in any case adapt to the new distribution of roles, re-discussing the question of *responsibility* in the interaction between provid-

---

[4] SPM – Society for Participatory Medicine: https://participatorymedicine.org/

[5] ACOR – Association of Cancer Online Resources: http://www.acor.org/

ers, insurers, industry, government regulation, artificial intelligence as well as patients, who are increasingly at the heart of things.

The question of responsibility also arises when, in future, smartphone and smart home devices provide advice not only in emergencies but also in the case of general medical issues. Amazon's Alexa digital home assistant, for example, already provides instructions on carrying out resuscitation [1].

19.5 Potential impacts on the healthcare economy

A key motivator for the transformation of medicine lies in the necessity of reducing the costs of healthcare provision. Industrialized nations use between 9% and 12% of their gross domestic product for covering healthcare costs. With healthcare spending equivalent to 17.8% of the GDP, the USA are particularly high in comparison with other nations [5].

19.5.1 Cost savings via objectified therapeutic decision-making

At around € 53 billion or 15.5% of overall spending, expenditure on medicines in Germany is a significant matter (cf. Fig. 19.9, [7]). In the USA, the sum amounts to more than $ 325 billion [5]. Just in 2010, cancerous conditions alone generated more than $ 120 billion in direct medical costs in the USA, with the trend rising sharply

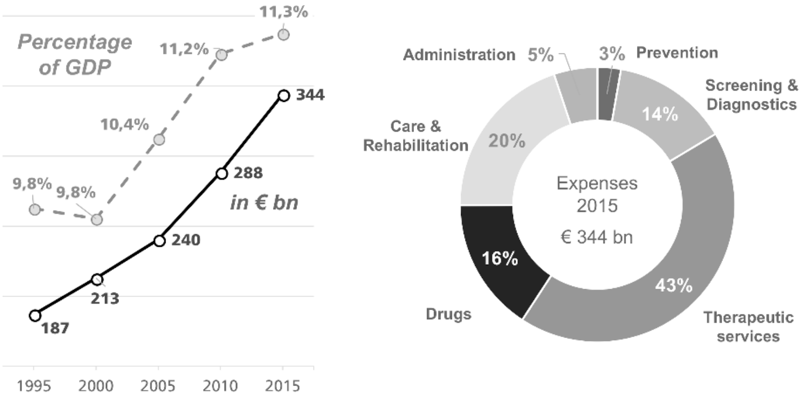


Fig. 19.9 Healthcare spending in the Federal Republic of Germany (Fraunhofer MEVIS, all rights reserved. Based on data from the Federal Statistical Office)

[15, 16]. Close to half of these costs are for medicines, in particular systemic chemotherapies and so-called "targeted therapies" that are focused more specifically on the disease in question and are available in increasing numbers. Typical annual therapy costs amount to between € 20,000 and € 100,000, and sometimes significantly more.

In many cases these expensive therapies remain unsuccessful or merely contribute to a minor delay in the progress of the disease before being discontinued. With the exception of specific types of cancer for which a therapy is possible, only around a quarter of drug-based cancer therapy treatments are seen to be effective (cf. [21]). Unsuccessful cancer therapies thus represent a considerable cost factor in the current healthcare system, at an annual cost of far more than $ 50 billion without providing a successful cure. What is worse for the patients is the fact that most chemotherapies have significant side effects that are only justified in the case of the intervention being successful.

Advances in specific immune therapy, in genomics as well as in early in vitro and in vivo therapy monitoring promise to provide a remedy via far more precise recommendations for the right combination of therapies, improved patient selection, and verification of therapy response [3].

19.5.2 Increasing efficiency via early detection and data management

In view of the distribution of tasks across the healthcare value chain (cf. Fig. 19.9), we see that by far the largest proportion of expenditure is on therapy, with an additional large proportion being allocated to care, and only a fraction remaining for prevention, early detection, and diagnosis. Since late diagnosis not only reduces chances of recovery but also increases the associated costs of therapy and care, early diagnostic detection offers significant potential savings. We still tend to diagnose too late and thus spend too much on therapy.

Digital patient records will also contribute towards cost efficiency in general patient management and the administration of medical services. Today's data silos and breakdowns in patient information sharing lead to a grave loss of information, in turn giving rise to redundant procedures in the form of additional medical consultations or diagnostic processes, for example.

19.6 Structural changes in the market

19.6.1 Disruptive innovation and the battle about standards

The vast majority of stakeholders in the healthcare system long for the digital transformation to take place. Due to extensive regulatory authorization requirements, tedious healthcare political processes surrounding design processes, and strong institutional integration, the healthcare system has thus far tended towards long technology lifecycles and incremental innovation. Disruptive innovation in medicine thus requires an alignment to the respective predominant national contexts. Disruption and cooperation need not stand in contradiction to one another. This is demonstrated by innovative healthcare solutions such as Oscar, an American healthcare insurance company reinventing the insurer as a mobile-first healthcare assistant, with a corresponding network of partner clinics already established. The considerable financial investments of American providers and insurers in innovative startups are also a witness thereof.

The vision of digitized medicine described above requires parallel innovations across the medical value chain. Simultaneously, the healthcare market is exhibiting characteristics of a network industry, tending, due to the effects of competition, towards the development of oligopolies with a small number of dominant market players. In the era of digitized medicine, comprehensive medically-curated patient data is of significant strategic value for the continuous optimization and validation of self-learning algorithms. In fact, we can assume that in the near future a standards battle about integrative authority in medicine and the associated access to patient data will arise. In order to safeguard the interoperability of the system and protect users from dependence on selected providers, it is desirable that standards be developed for sharing data in the same way as is already the case for clinic data (FHIR, HL7) and medical imaging data (DICOM).

19.6.2 New competitors in the healthcare market

The business magazine *The Economist* predicts that the transformation of medicine will lead to the development of a new competitive landscape, with established healthcare providers and pharmaceutical and medical technology giants on the one hand and technology insurgents on the other [24]. The latter include large technology companies such as Google/Alphabet, Apple, SAP, and Microsoft, but also a range of venture capital-financed startups, which are developing solutions for sections of digital medicine. Annual venture capital investments of more than $4 billion in the USA over the last three years demonstrate the hopes that investors are pinning on the healthcare transformation [23]. No less impressive is Google's latest funding commitment of $500 million for the above-mentioned Project Baseline.

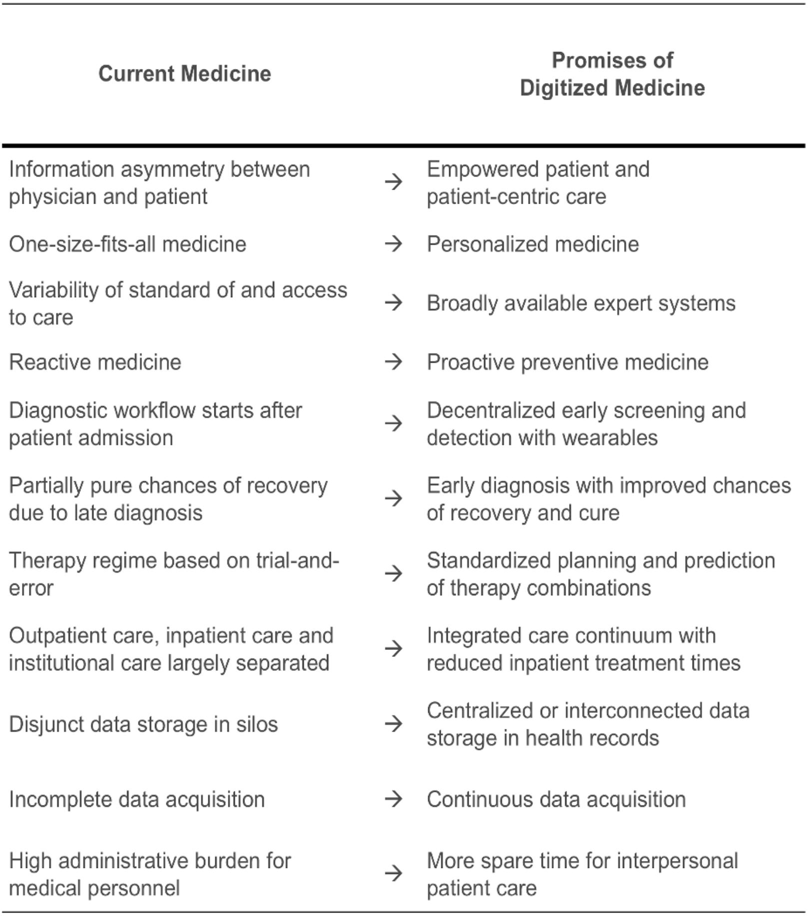
The decisive advantage of the technology insurgents with respect to established medical technology players, alongside their agility and abundant financial resources may, above all, be that in the innovation process they will neither be influenced by their own legacy nor by the danger of cannibalizing their existing business through disruption. Since AI-based software solutions will be a key differentiating feature in future competition, these firms have good prospects in the healthcare market. In the medium term, we expect a variety of partnerships to develop between providers, established players, and the new market participants, which will take on various forms according to national circumstances.

19.7 Outlook

Considering the digital transformation of medicine at some distance, we can see dynamic potential for growth with immense economic interest. A whole arsenal of opportunities for improvement is provided by the coming together of a vast range of technologies and developments, the so-called "super convergence". Some of these are the results of investments amounting to billions in biomedical basic research worldwide over recent decades. Of particular importance for the digital transformation will be the integration of hitherto unconnected medical data silos and the targeted use of the latest artificial intelligence methodologies. Consequently, objectified medicine will be created with structured treatment procedures and the effect of earlier and more accurate diagnosis with improved treatment outcomes and reduced costs (cf. Fig. 19.10).

Fraunhofer is a research and technology development partner for industry, politics, and clinics, and acts as a guide through the complexity of the digital transformation. Due to the multi-faceted nature of medicine and the constant growth in medical knowledge, lasting success can only be expected if the interlocking between technological and biomedical research on the one hand and clinical implementation and product development on the other is guaranteed long term. Population-focused research projects such as the German National Cohort will in future be even more closely interlinked with the findings of clinical data analysis in order to generate the best possible diagnosis and therapeutic recommendation.

The right combination of not-for-profit research, industrial development, clinical implementation, and smart political guidelines will also be key to deciding which



Current Medicine		Promises of Digitized Medicine
Information asymmetry between physician and patient	→	Empowered patient and patient-centric care
One-size-fits-all medicine	→	Personalized medicine
Variability of standard of and access to care	→	Broadly available expert systems
Reactive medicine	→	Proactive preventive medicine
Diagnostic workflow starts after patient admission	→	Decentralized early screening and detection with wearables
Partially pure chances of recovery due to late diagnosis	→	Early diagnosis with improved chances of recovery and cure
Therapy regime based on trial-and-error	→	Standardized planning and prediction of therapy combinations
Outpatient care, inpatient care and institutional care largely separated	→	Integrated care continuum with reduced inpatient treatment times
Disjunct data storage in silos	→	Centralized or interconnected data storage in health records
Incomplete data acquisition	→	Continuous data acquisition
High administrative burden for medical personnel	→	More spare time for interpersonal patient care

Fig. 19.10 The promises of digital medicine (Fraunhofer MEVIS)

advances will actually be achieved. Many healthcare provider have recognized the strategic and commercial value of their data and knowledge and are thus seeking the right route towards realizing this value. Therefore, prospective partnerships and business models need to also protect the legitimate interests of hospitals and patients so that a continuous exchange and build-up of data and knowledge can take place within the network. A well-known counterexample is the discontinued partnership between the MD Anderson Cancer Center and IBM Watson Health [12].

Alongside all of the legitimate optimism, particular efforts should be made to also recognize and account for the risks and dangers of data integration and automation during the design of future IT systems. In this regard, data protection and safety represents a critical concern and it is important to find the correct balance between protecting patients and facilitating medical progress, especially since a large proportion of patients would be prepared, according to surveys, to make their data available for research [25]. The approval processes for automated and self-learning software solutions, too, will require rethinking in the coming years. Last but not least, Fraunhofer is providing impetus for data security and digital sovereignty in Germany and Europe with its Big Data Alliance, for example, as well as the foundation of the Industrial Data Space Association[6]. The latter, founded at the beginning of 2016, currently counts around 70 institutions and firms among its members.

A further key issue will be adapting medical curricula to current advances. Doctors and nursing staff need to be prepared for the rapidly developing technological opportunities and needs of their patients. Even if we cannot predict the future accurately, care providers must nevertheless be in a position to orient themselves within the increasingly complex world of medical information and, especially when it comes to making use of enhanced computer assistance, make confident decisions. And let us not forget that empathy and human interaction are an essential contributing factor to successful recovery. Thus, the improvements in quality, security, and cost-efficiency of digital medicine should first and foremost enable nursing staff and doctors to spend more time with patients, freed from technical and bureaucratic burdens.

Sources and literature

[1] AHA–American Heart Association (2017) Alexa can tell you the steps for CPR, warn- ing signs of heart attack and stroke. Blog. Zugriff im Juli 2017: http://news.heart.org/ alexa-can-tell-you-the-steps-for-cpr-warning-signs-of-heart-attack-and-stroke/

[2] Aerts HJ, Velazquez ER, Leijenaar RT et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006. doi:10.1038/ncomms5006

[3] ASCO (2017) Clinical Cancer Advances 2017, American Society of Clinical Oncology. Zugriff im Juli 2017: https://www.asco.org/research-progress/reports-studies/clinical-cancer-advances

---

[6] Industrial Data Space Association: http://www.industrialdataspace.org

[4] CB Insights (2017) From Virtual Nurses To Drug Discovery: 106 Artificial Intelligence Startups In Healthcare. Zugriff im Juli 2017: https://www.cbinsights.com/blog/artificial-intelligence-startups-healthcare/

[5] CMS – Centers for Medicare and Medicaid Services (2017) NHE Fact Sheet. Zugriff im Juli 2017: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html

[6] Cooper DN, Ball EV, Stenson PD et al (2017) HGMD – The Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff. Zugriff im Juli 2017: http://www.hgmd.cf.ac.uk/

[7] Destatis – Statistisches Bundesamt (2017) Gesundheitsausgaben der Bundesrepublik Deutschland. Zugriff im Juli 2017: https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/Gesundheitsausgaben/Gesundheitsausgaben.html

[8] Dusheck J (2016) Diagnose this – A health-care revolution in the making. Stanford Medicine Journal, Fall 2016. Zugriff im Juli 2017: https://stanmed.stanford.edu/2016fall/the-future-of-health-care-diagnostics.html

[9] Esteva A, Kuprel B, Novoa RA et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118. doi:10.1038/nature21056

[10] Ferrucci D, Levas A, Bagchi S et al (2013) Watson: Beyond Jeopardy! Artificial Intelligence 199:93–105. doi:10.1016/j.artint.2012.06.009

[11] Harz M (2017) Cancer, Computers, and Complexity: Decision Making for the Patient. European Review 25(1):96–106. doi:10.1017/S106279871600048X

[12] Herper M (2017) MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine. Forbes. Zugriff im Juli 2017: https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine

[13] Knight W (2017) The Dark Secret at the Heart of AI. MIT Technology Review. Zugriff im Juli 2017: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

[14] Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48(4):441–6. doi:10.1016/j.ejca.2011.11.036

[15] Mariotto AB, Yabroff KR, Shao Y et al (2011) Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst 103(2):117–28. doi:10.1093/jnci/djq495

[16] NIH – National Institutes of Health (2011) Cancer costs projected to reach at least $158 billion in 2020. News Releases. Zugriff im Juli 2017: https://www.nih.gov/news-events/news-releases/cancer-costs-projected-reach-least-158-billion-2020

[17] Ryan KJ (2016) Who's Smartest: Alexa, Siri, and or Google Now? Inc. Zugriff im Juli 2017: https://www.inc.com/kevin-j-ryan/internet-trends-7-most-accurate-word-recognition-platforms.html

[18] Sahiner B, Chan HP, Petrick N et al (1996) Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 15(5):598–610. doi:10.1109/42.538937

[19] Schmutzler R, Huster S, Wasem J, Dabrock P (2015) Risikoprädiktion: Vom Umgang mit dem Krankheitsrisiko. Dtsch Arztebl 112(20): A-910–3

[20] Silver D, Huang A, Maddison CJ et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489. doi:10.1038/nature16961

[21] Spear BB, Heath-Chiozzi M, Huff J (2001) Clinical application of pharmacogenetics. Trends Mol Med 7(5):201–4. doi:10.1016/S1471-4914(01)01986-4

[22] Stenson et al. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next- generation sequencing studies. Hum Genet 136:665-677. doi: 10.1007/s00439-017- 1779-6

[23] Tecco H (2017) 2016 Year End Funding Report: A reality check for digital health. Rock Health Funding Database. Zugriff im Juli 2017: https://rockhealth.com/reports/2016-year-end-funding-report-a-reality-check-for-digital-health/

[24] The Economist (2017) A digital revolution in healthcare is speeding up. Zugriff im Juli 2017: https://www.economist.com/news/business/21717990-telemedicine-predictive-diagnostics-wearable-sensors-and-host-new-apps-will-transform-how

[25] TheStreet (2013) What Information Are We Willing To Share To Improve Healthcare? Intel Healthcare Innovation Barometer. Zugriff im Juli 2017: https://www.thestreet.com/story/12143671/3/what-information-are-we-willing-to-share-to-improve-healthcare-graphic-business-wire.html

[26] Topol E (2012) The Creative Destruction of Medicine: How the Digital Revolution will Create Better Health Care. Basic Books, New York. ISBN:978-0465061839

[27] Trotter F, Uhlman D (2011) Hacking Healthcare – A Guide to Standards, Workflows, and Meaningful Use. O'Reilly Media, Sebastopol. ISBN:978-1449305024

[28] Zhang W, Hasegawa A, Itoh K, Ichioka Y (1991) Image processing of human cor- neal endothelium based on a learning network. Appl Opt. 30(29):4211–7. doi:10.1364/AO.30.004211

Smart Energy

20

The digital transformation in the energy sector

Prof. D. eng. Peter Liggesmeyer ·
Prof. Dr. Dr. h.c. Dieter Rombach · Prof. Dr. Frank Bomarius
Fraunhofer Institute for Experimental Software
Engineering IESE

Summary

A successful energy transition is inconceivable without extensive digitization. In view of the complexity of the task of digitizing the energy sector and all of the associated systems, previous efforts at defining essential components of digitization (such as concretely usable reference architectures and research into the resilience of the future energy system) currently still appear insufficient and uncoordinated. These components include smart management approaches capable of integrating market mechanisms with traditional management technologies, and comprehensive security concepts (including effective data utilization control) that need to go far beyond the BSI security profile for smart meters. The digitization of the energy system needs to be conceived and operated as a transformation process designed and planned for the long term with reliable milestones.

20.1 Introduction: The digital transformation megatrend

Digitization facilitates the smart networking of people, machines, and resources, the ongoing automation and autonomization of processes, the personalization of services and products, as well as the flexibilization and fragmentation, but also the integration of business models across the entire value chain [8]. In the context of this definition, digitization is increasingly understood as a process of transformation, which gives rise to the opportunity to scrutinize processes and procedures fundamentally and align them to revised or often even completely new business models.

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_20

In order to achieve this, standard system architectures in most cases need to either be fundamentally revised or completely recreated so that conventional, modernized, and new products and services can be offered via new network topologies and communications. The term "smart ecosystem"[1] was coined for these kinds of new business models developed during the course of digitization. The fundamental characteristic of the smart ecosystem is the interplay between individual business models within a larger overall system oriented towards economic goals. The information technology basis for this is formed by standards and open IT platforms with low transaction costs, high reliability, and security for technically implementing the business models. Internet technologies provide the universal infrastructures for connecting business partners with one another as well as linking the digital representations of things, devices, installations, goods, and services (so-called "digital twins") with the physical objects accessible via the Internet (Internet of Things, IoT).

This transformation has already led, in numerous domains, to business models previously linked to physical objects and personally provided services being superimposed or even substituted by a dematerialized data economy (hybrid value creation) in the course of digital transformation. Digitally provided services are thus increasingly in the foreground:

- Amazon delivers without being a producer, Uber provides transport without having its own fleet, Airbnb lets rooms without owning accommodation.
- Digitization avoids unprofitable downtime of assets through predictive maintenance.
- Digitization replaces investments in facilities with service rental (logistics, heating, lighting, etc.)
- Digitization facilitates new contract models, e.g., on-time arrival contracts for train journeys (Siemens)
- Digitization adds value to traditional products (John Deere FarmSight)

Key developer and operator competencies within a data economy are cloud and IoT technologies, (big) data analytics, machine learning, and deep learning.

The act of fundamentally revising business processes frequently comes up against the limits of applicable legal frameworks. Data protection and the protection of privacy take on new meaning during the process of intensive digital networking and need to be revised in order to guarantee sufficient protection on the one hand,

---

[1] A smart ecosystem integrates information systems that support business objectives and embedded systems that serve technical objectives so that they operate as one and can pursue shared overarching (business) objectives.

and to permit new and economically sustainable business models on the other hand. Alongside economic reasons, the (current) lack of an internationally harmonized legal framework may lead to the relocation of firms to countries with more accommodating legislation.

In highly regulated fields such as healthcare, food products, transport, and the energy sector, legal regulations and compulsory standards such as the EU's General Data Protection Regulation[2] should be taken into account in good time and, where necessary, should be revised in order to not delay the transformation process unnecessarily or even smother it completely.

20.2 Digital transformation in the energy sector

The digital transformation in the energy sector is a process that will unfold relatively slowly over time, and will most likely require several decades. This is due to the following factors:

- The long-term nature of investment decisions regarding extremely expensive network infrastructure and power plants requires decision-making that provides economic security, which is particularly difficult against the backdrop of the far-reaching structural change provoked by the energy transition. It must nevertheless be ensured that this aspect is not misused as a reason for unnecessary delays in digitization, which is needed so urgently.
- The weight of regulation in those areas of energy supply that operate as a natural monopoly limits innovation since the regulatory adjustments necessary for operating new equipment and business models generally lag behind the innovation itself.
- The transformation process affects several sectors of the energy industry that have thus far operated separately (electricity, gas, heating) and impacts associated sectors (transport, home automation, industrial automation). Due to the dominance of the renewable energy sources of wind and sun and their volatile feed-in, the electricity sector represents the controlling variable here to which the other sectors must adjust.

In the course of the process of digital transformation, traditional energy sector products will fade into the background and be replaced by other services with cash value qualities. Formerly, products such as electricity, heating, and gas were primar-

---

[2] EU General Data Protection Regulation (GDPR, EU Regulation 2016/679), effective May 25, 2018

ily invoiced according to the physical unit of work (in multiples of watt-hours, Wh), that is, according to a volume tariff, or sometimes according to the maximum supply availability (in multiples of Watts, W). The quality of the energy supply from centrally managed, high-capacity power plants was usually very high but was not an explicitly stated component of pricing. In other words, the costs of security of supply and power quality[3] have thus far been priced into the volume tariff.

The development of renewable sources of energy and the politically formulated climate goals both result in the dismantling of power plants (nuclear phase-out and planned phase-out of coal-fired power plants for decarbonization purposes) within the electrical energy grid. Centrally provided security of supply, network services[4], and power quality will thus be lost. In the future, they must be provided and invoiced by the renewable energy feeder operating in a decentralized manner, and will thus gain the significance of tradable products.

Additional new products could, for example, include fixed levels of service even during critical conditions within the energy network, similar to the Service Level Agreements (SLAs) in the information and communications (ICT) sector. For industrial customers, scattered examples of such products already exist; in the course of digitization, they may become standard, ensuring for example that basic, low-energy functions are maintained and dispensable devices are selectively turned off during periods of extreme electrical energy undersupply, where until now complete subnetworks had to be shut down. Developing this thought further means that instead of paying for used, fed-in, or transported energy, **flexible markets** will develop in the future where changes in the feed-in, requirements, or consumption profiles can be traded, managed, and invoiced in cash value as flexibly as possible and calculated by digital systems. Flexibility is above all a product motivated by the technical necessities of operating a system with highly fluctuating renewable energy feed-in levels. Essentially, it focuses on the costs of operating under unbalanced network conditions and only prices the actual amount of energy flow indirectly, particularly since the marginal costs of renewable energy plants tend towards zero.

In the future then, energy quantities or power ratings will rather play the role of unchangeable physical limitations for the digital markets. Even today, the actual

---

[3] In simple terms, **power quality** refers to the maintenance of the target voltage (230V) and the target frequency (50Hz). The actual voltage may thus deviate by 10 percent and the frequency by 0.2 Hz. Greater deviation may entail damage in connected systems (both producers and consumers), and power supplies are thus generally subjected to emergency shutdown.

[4] **Network services** refer to technically necessary functions, e.g., reactive power generation, black-start capabilities (restarting after a blackout), and provision of short-circuit power (maintaining the flow of electricity in the case of a short circuit such that fuses can trigger).

significance of the demand and consumption charges is already greatly reduced even though invoicing is still based on consumption and demand. Occasionally, there are discussions regarding the kinds of flat rates common in telecommunications. But the current energy mix, efficiency demands, and climate goals still stand in the way of flat rates for energy.

Before digitization can really transform the energy sector in the manner described above, however, distribution grids must be extensively equipped with information and communications technology on the one hand, and new roles must be defined for the operators of this technology within a data economy on the other hand.

Table 20.1 Theses on digitization in the energy system of the future

1	Necessary key technologies such as the Internet of Things and Industry 4.0 are either already available or will soon be usable (5G).
2	Energy flows are increasingly accompanied by information flows – smart meters are just the beginning. Similar to the manufacturing sector, "digital energy twins" and an "energy data space" [14] are developing.
3	Energy-related data is becoming a valuable asset – an energy data economy is developing.
4	The significance of sector coupling and of markets is growing. Interactions between previously separate systems (e.g., electricity, gas, heating, e-mobility) are being developed in the process; digital ecosystems are coming into being that are tightly networked with one another.
5	The digitized energy system is trustworthy and behaves as expected.
6	Self-learning, adaptive structures support ongoing planned as well as erratic changes in the energy system.
7	The energy system of 2050 will have significantly higher resilience requirements due to its decentralized and heterogeneous nature; at the same time, decentralization and heterogeneity are also part of the solution to the challenge of resilience.

20.3 The energy transition requires sector coupling and ICT

Ambitious climate goals can only be achieved via massive decarbonization. The German approach of largely replacing fossil fuels while simultaneously phasing out nuclear energy requires a massive expansion of renewable energy installations for years to come. Wind and sun are available in Germany, but water power only to a limited degree. Biogases are only able to supply limited quantities of energy, especially due to the emerging competition with the food production industry. Thus, supply based on sun (PV and solar heat) and wind is extremely dependent on the weather and the time of day – in other words, we have volatile or fluctuating feed-in. Fluctuation – which may mean both a dramatic temporary oversupply of electrical energy and a massive undersupply (volatility) that may last for days – has to be handled appropriately, without reducing the usual security of supply. A portfolio of measures and goals for solving the problem is under discussion:

- **Efficiency measures** for reducing the energy requirements of devices and installations:
  While measures to increase efficiency are basically to be supported since they reduce energy costs and make it easier to achieve climate goals, they make very little contribution to solving the problem of fluctuating provision of electrical energy. Energy efficiency can also be increased by means of non-digital technical improvements, for example by reducing heat losses, pressure losses, or friction. In, numerous modern devices, however, smart digital controls are responsible for increased energy efficiency, for example in the case of smart drive control systems in pumps. The design of these control systems has thus far been oriented towards optimization of their operation in the face of energy costs that are neither time- nor load-dependent. It is only by opening the systems' communications to signals from the supply network (e.g., coded as variable tariffs) that these control systems can also become useful to the network. This is termed "demand-side management".

- **Adapting consumption to supply** (demand-side management, DSM, for load-shifting and for activating storage capacity in end-customer installations):
  The energy requirements of devices and installations can (also) be regulated with a view to the current grid situation – e.g., by changing load profiles or utilizing storage capacity in the devices and installations in order to influence the period/amount of energy use or energy provision. This flexibility potential can then be used to compensate for oversupply or undersupply in the electricity grid. Only through extensive digitization and communicative interconnection of both the energy networks and the feed-in and consumer installations can this kind of network-supporting behavior via DSM be achieved.

- **Sector coupling** in order to make the flexibility potential of other energy systems (gas, heating, cooling) usable for the electricity sector:

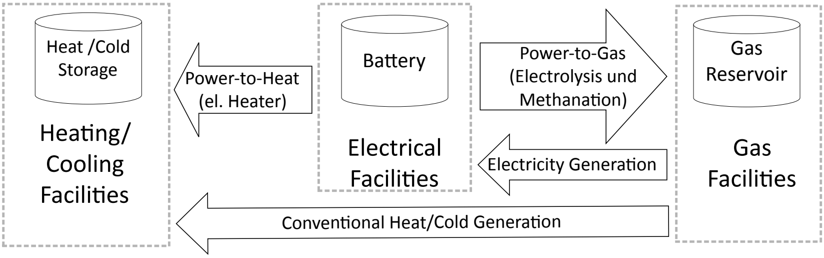


Fig. 20.1 Sector coupling of heating, electricity, and gas (Fraunhofer IESE)

In sector coupling, flexibility potentials are harnessed in a way that goes beyond the possibilities of DSM: either alternative energy sources are momentarily used to cover energy demand in the face of a shortfall of electrical energy (e.g., by using gas instead of electricity), or the storage capacity of another form of energy (heating, cooling, etc.) is used to adjust the load profile on the electricity side, which is generally cheaper than using a battery (see Fig. 20.1). Sector coupling is considered to have the greatest potential – in terms of quantity and cost – for providing the necessary flexibility for the electricity grid of the future. It can be used in both small end-customer installations, such as in the kW domain as bivalent heating using a choice of either gas or electricity, and in the MW domain, e.g., for energy suppliers providing additional electric heating for heat storage in district heating networks. If there are significant excess supplies of electricity, variants of sector coupling with lower levels of effectiveness may also be economically feasible, such as power to gas (electrolysis, methanation) or power to liquid (production of liquid fuels) [5].

- **Market mechanisms:**
  Flexibility requests arising from volatile feed-ins or imbalances in the system and flexibility offerings resulting from demand-side management and sector coupling may in the future be traded on new electronic markets in order to achieve a balanced power economy within the electricity network, primarily through the use of market mechanisms. Market mechanisms are favored politically, and the quest for new business models that identify and test out these mechanisms is under way. A number of research projects to this end are investigating the potential interaction of technical, market-related, and adapted regulatory conditions (e.g., [9][15][16]).

With the exception of energy efficiency, the four approaches described above can thus be implemented exclusively by using digitization: Devices and installations connected to the electrical energy system need to be digitalized and networked communicatively in order to identify, communicate, negotiate, and ultimately measure and invoice flexibility potentials offered and requested. The term "**smart energy**", frequently used in this context, refers to the change from centralized control based on predicted consumption to decentralized control, where the actual offerings and requests are balanced out regionally as far as possible, using market mechanisms in real-time. Only if these market mechanisms fail will it be necessary to revert to controlled measures (control loops), which will then override these market mechanisms temporarily and regionally. The energy system of the future must thus have a range of automated control strategies at its disposal to be utilized according to the situation at hand. Significant research work is still required in order to define and test these strategies with a view to ultimately guaranteeing resilient operation. In order to accommodate decentralized feed-in and control, the traditional centralized hierarchical arrangement of electricity grids must be replaced by a cellular hierarchical arrangement.

20.4 The cellular organizational principle

Geographical imbalances in generation and consumption ultimately need to be balanced out via electrical grids. Extreme imbalances (regarding amount of energy, power, and local distribution) require correspondingly powerful and thus expensive grids.

For historical reasons, the topology of today's electrical grids is organized hierarchically and is split into various levels of voltage. While large distances are bridged with voltages in the region of several hundreds of thousands of volts (transmission grids), this voltage is reduced via transformers across several grid levels (medium-voltage grids) until it finally reaches the 230/400 volts (distribution grids) required for residential connections.

In the past, energy was fed-in at the highest voltage levels by large power stations. From there, the electrical energy was distributed to the wider region. The flow of energy within the wiring and the transformers was unidirectional. Security of supply, securing of the power quality, and provision of grid services primarily took place via interventions at the high-voltage levels. The hierarchical grid structure was tailored to this centrally controlled generation by means of small numbers of large power stations (operated with nuclear or fossil fuels).

In the course of the energy transition, the number of these large power stations will decrease, while at the same time more and more energy will be fed-in to the middle and lower voltage levels from renewable energy installations (decentralized

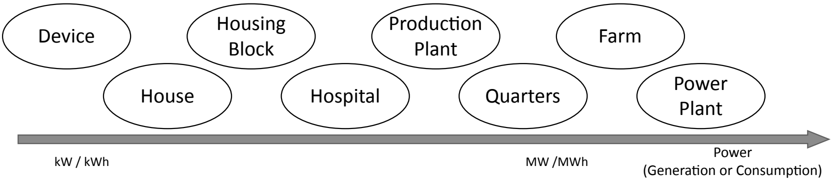


Fig. 20.2 Examples of cell sizes (Fraunhofer IESE)

feed-in). Energy will have to be able to be transported bidirectionally since overload caused by temporary local excess feed-in from renewable energy installations will have to be transferred from their grid section to the higher voltage levels for onward transmission. Wiring and transformers would need to be upgraded for this purpose, or the generating installations would have to be deactivated temporarily. Sector coupling and DSM could also make it possible to temporarily raise local consumption in a targeted manner in the respective grid section. A similar process applies for insufficient feed-in within the grid section.

Provided that regional generation and consumption are approximately balanced in the case of decentralized feed-in, the grid load will decrease. Sector coupling and DSM provide effective levers here. However, for cost reasons these regional “cells” cannot be completely independent. In addition, we have differing geographical concentrations of wind (northern Germany) and sun (southern Germany); electrical transmission and distribution grids will thus not become obsolete [4]. Digitization may, however, very well limit the grid load or the necessity for development, at least at the distribution grid levels.

Fig. 20.2 shows examples of potential cell sizes arranged along an informal scale. The cell concept is a recursive concept. This means that higher-level cells (e.g., districts) may be formed of lower-level ones.

The challenges of decentralized feed-in suggest **cellular** control structures with **subsidiary hierarchical** distribution of roles. Sector coupling and DSM take place within the cells. Cells are able to achieve energetic balance between one another and within the hierarchy (see Fig. 20.3).

Cellularity is also a requirement from a system security point of view: centralized systems generally have a single point of failure. As soon as an essential (non-redundant) system component fails or is compromised by a physical attack or a cyber-attack, the system as a whole is no longer capable of functioning. Systems operating in a decentralized manner are harder to attack due to the greater number of components that have a role in operations, and the attack’s reach is, in the first instance, limited to the components attacked or their respective cells. Cells can thus limit the spread of

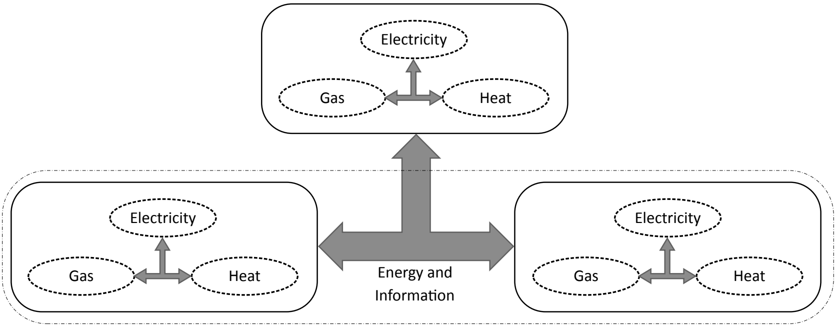


Fig. 20.3 Schematic diagram of a hierarchical cellular structure with sector coupling (Fraunhofer IESE)

undesirable grid conditions. Nevertheless, in the age of the Internet and automated attacks, this advantage may be lost if the architecture of the cell control systems does not contain suitable lines of defense, or if the ICT in the cells is identically implemented and thus any weak points disclosed can be attacked on a broad scale with limited effort. **Diversity**[5] of software implementation instead of digital monocultures in control rooms, operating systems, and control algorithms makes broad-scale attacks more difficult. Diversity is thus a desirable – if costly – system property, since economies of scale are lost and any solutions found are more difficult to transfer.

Digitization makes it possible to utilize flexibility as a “currency” within the energy system of the future. With flexibility as a product, new roles and actors will be defined in the course of digital transformation, and new business models will arise. The sector is currently beginning to think about which new personalized products could be conceivable and profitable. A few examples are:

- Software plus consultancy services for planning and designing cell infrastructure (for existing stock and new plans).
- Software for continuously monitoring cells vis-à-vis a range of indicators, e.g., climate protection contributions, energy flows, material flows, logistics, security status. Services around the generation, analysis, and distribution of this data.
- Various control rooms for actors at the cell level (aggregators, contractors, data resellers) with corresponding data aggregation and processing functions.

---

[5] Diversity in implementation is not in contradiction to urgently needed standardized interfaces and exchange formats. The latter are indispensable for efficient operation. The implementations “behind” the interfaces, however, should differ, as viruses and Trojan horses are written for the most common types of software.

- Measurement and analysis software for end customers (prosumers), landlords, energy suppliers, aggregators, and other roles with corresponding services.
- Analysis software for identifying and implementing value-added services for suppliers, aggregators, and end customers.

20.5 Challenges for energy ICT

In recent years, the information and communications sector has clearly demonstrated that large ICT systems can be built and operated to be highly reliable, powerful, and profitable. Leading examples are the Internet or the global mobile telephony systems. In this respect, the digitization in the course of the energy transition – involving several hundred million devices and installations across Germany in the long term, or billions worldwide – is no more demanding in terms of numbers and ICT performance than global mobile telephony, electronic cash, or IPTV.

Effective mechanisms for data security and data usage control are essentially well-known. The issue now is to consistently build these into the resulting systems right from the start instead of having to upgrade them later.

With significant participation from Fraunhofer, current German flagship projects are demonstrating how industry standards have been systematically created for years in the area of industry-relevant embedded systems in order to advance comprehensive digitization. Examples include AUTOSAR [9], BaSys 4.0 [10], or the Industrial Data Space (IDS) [11]. Flagship projects for the energy systems sector with comparably high demands are just starting: see projects within the programs SINTEG [15] and KOPERNIKUS [16]. The criticality for society and industry of the ICT systems of the future energy infrastructure will be key to their design. Based on the high availability and power quality that is customary in Germany, the future energy system, which will be several orders of magnitude more complex[6], should deliver at least the same availability and power quality.

Every system accessible via the Internet nowadays is exposed to a stream of increasingly sophisticated cyber-attacks, which is growing by far more than a factor of ten every year [18]. The energy system of the future, too, will be based on Internet technology and will be a valuable target for attacks from the Internet. Despite all precautionary safeguards and redundancy, breakdowns due to energy installation

---

[6] In Germany, more than 560,000 transformers and more than 1.5 million renewable energy installations were already part of the electricity grid as of 2015 [17], but there is hardly any communication among them yet. A hypothetical area-wide smart meter roll-out would add around 40 million networkable meters, not counting controllable consumers and batteries.

damage, extreme operating conditions, and breakdowns in the electrical grid, as well as due to the breakdown of communication networks and faulty behavior due to targeted physical and cyber-attacks on installations and grids are all to be expected. In particular, these situations may occur in complex combinations or may be provoked in a targeted manner. There is thus absolutely no doubt that breakdowns will occur in a system of this complexity. The attack vectors are becoming ever more complex and more difficult to recognize.

In addition, the energy system of the future will be configured significantly less statically than the present system. With the huge number of generating and consumer installations, physical additions to and disconnections of installations will, from a statistical point of view, occur far more frequently. Not to mention the fact that when market participants are permitted to act freely, their affiliations with service provider balancing groups will change (virtually). Electromobility – particular the use of fast-charging systems – equally implies a temporary reconfiguration within the network, with the end customer probably wanting to be assigned to their home energy supplier while traveling (roaming).

In the context of a constantly changing system and simultaneously high requirements for security of supply as well as for safety and security, the conventional "fail-safe" design principle[7] is no longer adequate. The system instead needs to be "safe to fail" [6]. This means that even when significant components break down or their performance degrades, the remaining system automatically responds to the situation without breaking down completely[8]. Mechanisms for achieving this include, for example, runtime adaptations or optimization; a range of "self-x" technologies are thus indispensable for the energy system of the future:

- **Self-diagnosis:**
  The system needs to constantly monitor essential system parameters and indicators in order to assess its current condition with respect to security, stability, and reserves.
- **Self-organization** (adaptation, self-healing):
  As soon as important system parameters are in danger of becoming critical or have already become critical, the system must assess alternative configurations and move towards more stable and secure conditions via reconfiguration or a changed behavior profile.
- **Self-learning:**
  Systems as complex as the energy system cannot be programmed or configured manually. Even the most minor changes would entail unreasonable overhead.

---

[7] The system no longer works, but transitions to a safe state in any case.

[8] In the automotive sector, this operational state is called fail-operational.

The system itself needs to capture its condition, connected installations, as well as their parameters and typical profiles, map them to models, and actively use the results for control purposes (model learning).

- **Self-optimization:**
  A cell's many and diverse optimization goals may change dynamically. Climate protection goals, for example, may vary in their importance according to the time of day or year. The system needs to be able to adapt itself to optimization goals, which may in part be contradictory and vary according to different timescales.

The foundation for fulfilling all of these requirements is the massive collection, processing, and secure sharing of data. To do this, an **Energy Data Space** (EDS) must be created as an adaptation of the Industrial Data Space (IDS) [12].

A fundamental ICT **reference architecture** needs to be defined that, in particular, specifies and permits the implementation of all of the key requirements with respect to security (with the goal of being safe to fail), sets standards, but does not stand in the way of a multitude of implementations (goal: diversity).

Whereas in other sectors overarching ICT standards are being created in a targeted manner, e.g., via AUTOSAR [9] for automotive engineering or BaSys 4.0 [10] and IDS [11] for the field of embedded systems and Industry 4.0, the ICT landscape in the energy field thus far still resembles more of a patchwork rug.

20.6 The challenge of resilience and comprehensive security

The energy transition is characterized by decentralization, both regarding the generation of energy and the control of the energy system, by volatile supply and by massive digitization. Set against the backdrop of this far-reaching upheaval, a suitable concept of resilience needs to be defined and operationalized for the energy sector [13]. In order to achieve this, and in order to make fundamental decisions on the system design, the responsibility for the provision of system support services in this context must be redistributed. These services include, for example, maintaining voltage and frequency, supplying reactive power, providing secondary operating reserves, tertiary control, black-start support, short-circuit power, and also, where necessary, primary operating reserves (cf. also "The Role of ICT" in [7]). How much decentralized ICT is actually necessary and sensible for which tasks also needs to be defined primarily against the backdrop of resilience and real-time requirements.

As explained in the previous section, the future energy system will not be configured statically and will need to assert itself in the face of changing and at times unexpected influences and attacks.

Resilience is the ability to adapt to previously unknown and unexpected changes and simultaneously continue to provide system functions.

Modern definitions of the term "resilience" refer to its close relationship to the concepts of security, forecasting, and sustainability [2][6]. To this extent, resilience is not a schematic response to negative influences but also incorporates the ability to self-adapt and to learn.

Traditional risk analyses are reaching their limits due to the complexity of the energy system. They are also only suitable to a limited degree for identifying new and unexpected events. In the future, criteria will therefore be required for operationalization and quantification of the resilience of the energy system during operation. In some cases, however, the criteria, methods, and indicators for measuring resilience first need to be developed. Monitoring technologies need to be combined with a systemic approach in order to identify the energy system's potential vulnerabilities already during its transformation (i.e., without interruption) [3]. Functioning examples of how system security and functionality can be monitored and ensured during operation – even in the face of changes to the system configuration – are well known from other sectors (Industry 4.0 or commercial vehicles).

In the energy system of the future, a previously unknown interaction between physical and virtual elements will develop. As a result, suitable new strategies for redundancy also need to be worked out. The tried-and-tested "n-1" rule for redundancy design will thus be insufficient to compensate for all of the diverse potentially erroneous interactions between ICT systems (which are corruptible), potentially maliciously influenced markets, regulated subsystems, and unchangeable physical constraints. The property of being "safe to fail" always extends across all of the system's physical and virtual components – resilience is a property arising from complex interactions that is specifically organized in each cell based on the individual configuration of the installations within a cell. Suitable early warning systems and the implementation of elastic system reactions are pressing research questions.

Closely related to the issue of resilience are issues surrounding the forecasting of system behavior, the immediate observability of the system state, and the reliable transparency of system actions. How this kind of monitoring system should be constructed and how complex system states and complex interrelated processes can be represented in a way that is clear and understandable for the user is yet another important research question.

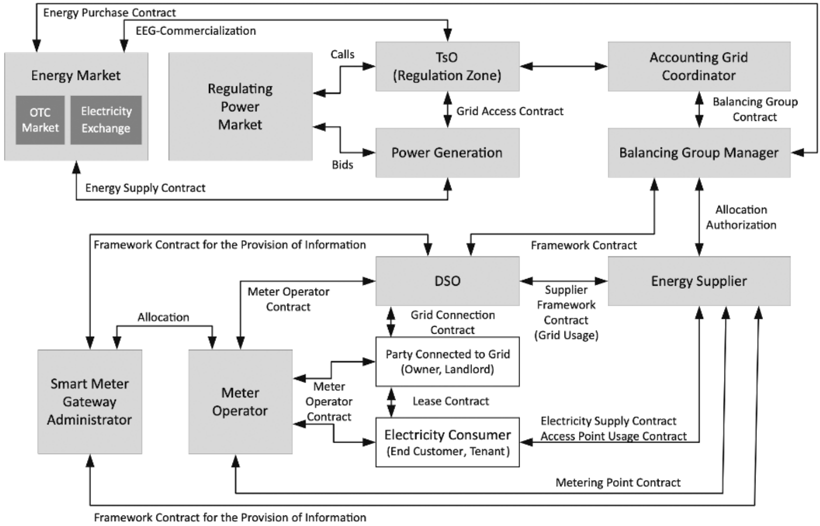


Fig. 20.4 Present contractual relationships between energy sector players (Bretschneider, Fraunhofer IOSB-AST)

Finally, set requirements and the services provided must be documented comprehensively and in a way that is subject to neither falsification nor dispute, and must be made accessible to invoicing. Until now, this has been regulated via a complex, static web of contracts (cf. Fig. 20.4) and corresponding reporting channels stipulated by the BNetzA (Federal Network Agency).

Consistent establishment of market mechanisms throughout the advancing energy transition will lead to very small volumes of energy (resp. flexibilities) being traded and the contract partners here (mainly “prosumers”) not being able to first conclude bilateral master contracts. The contractual agreements that are necessary to protect the large volume of brief relationships between the numerous actors within cells require contracts concluded by machine (“smart contracts”) that are founded on framework agreements concluded by humans. Here, too, there is a need for research in order to conclude legally protected agreements in real time between machines that must then be translated into system actions, monitored during execution, traceably documented, and correctly invoiced. At the moment, the blockchain approach is being propagated in this context. Its suitability remains to be verified.

20.7 The energy transition as a transformation process

The energy transition is a process that will require several decades due to its technical and societal complexity. During the course of the transformation process, old and new technologies will need to not only coexist but function in an integrated manner over a long period of time. The authors are convinced that now is the time to focus more intensively on the digitization of the energy transition. Only when the energy transition is understood as a complex and systemic transformation process can digitization actively support and successfully shape the necessary changes at the technical and societal level and help to press ahead with the transformation process. Very detailed support for this assessment is provided by the Münchner Kreis in its 50 recommendations [1].

In the past, there were many important innovations with respect to renewable energy technology. The accompanying digital networking and the resulting systemic challenges and opportunities have been neglected for a long time. Although it appears necessary and has often been discussed, the cellular approach has thus far been the focus of far too little research in the context of the energy system. The specification and implementation of resilience – ultimately a critical system characteristic affecting all of its components – also remain largely unexplored.

Last but not least, the success of the technical transformation process is, to a very large extent, reliant on long-term social acceptance and support [19], which must be continually verified and actively designed.

Sources and literature

[1] Münchner Kreis: 50 Empfehlungen für eine erfolgreiche Energiewende, Positionspapier, 2015.

[2] Neugebauer, R.; Beyerer, J.; Martini, P. (Hrsg.): Fortführung der zivilen Sicherheitsforschung, Positionspapier der Fraunhofer-Gesellschaft, München, 2016.

[3] Renn, O.: Das Energiesystem resilient gestalten, Nationale Akademie der Wissenschaften Leopoldina, acatech – Deutsche Akademie der Technikwissenschaften, Union der deutschen Akademien der Wissenschaften, Prof. Dr. Ortwin Renn Institute for Advanced Sustainability Studies, 03. Februar 2017 ESYS-Konferenz.

[4] Verband der Elektrotechnik, Elektronik, Informationstechnik e. V. (Hrsg.): Der zellulare Ansatz – Grundlage einer erfolgreichen, regionenübergreifenden Energiewende, Studie der Energietechnischen Gesellschaft im VDE (ETG), Frankfurt a. M., Juni 2015.

[5] Dötsch, Chr.; Clees, T.: Systemansätze und -komponenten für cross-sektorale Netze, in: Doleski, O. D. (Hrsg.): Herausforderung Utility 4.0, Springer Fachmedien GmbH, Wiesbaden, 2017.

[6] Ahern, J.: From fail-safe to safe-to-fail: Sustainability and resilience in the new urban world, Landscape and Urban Planning 100, 2011, S. 341–343.
[7] H.-J. Appelrath, H. Kagermann, C Mayer, Future Energy Grid – Migrationspfade ins Internet der Energie, acatech-Studie, ISBN: 978-3-642-27863-1,2012.
[8] Bundesnetzagentur: Digitale Transformation in den Netzsektoren – Aktuelle Entwicklungen und regulatorische Herausforderungen, Mai 2017.
[9] Flex4Energy FLEXIBILITÄTS¬MANAGEMENT FÜR DIE ENERGIE-VERSORGUNG DER ZUKUNFT, https://www.flex4energy.de
[10] AUTOSAR (AUTomotive Open SystemARchitecture) https://www.autosar.org/
[11] Basissystem Industrie 4.0 (BaSys 4.0). http://www.basys40.de/
[12] White Paper Industrial Data Space https://www.google.de/url?sa=t&rct=j&q=&e src =s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwiHp6X9qr7UAhU Fl xoKHeBcAFUQFggvMAE&url=https%3A%2F%2Fwww.fraunhofer.de%2Fc- on tent%2Fdam%2Fzv%2Fde%2FForschungsfelder%2Findustrial-data-space%2FIndu strial-Data-Space_whitepaper.pdf&usg=AFQjCNGHaxIr7RxDWl6OPIaUZlwGaN0R ZA&sig2=_mec3o8JzqMQDxLeQTkldw
[13] Dossier der Expertengruppe Intelligente Energienetze: Dezentralisierung der Energienetzführung mittels IKT unterstützen, Juni 2017
[14] Dossier der Expertengruppe Intelligente Energienetze: Branchenübergreifende IKT-Standards einführen, Juni 2017
[15] https://www.bmwi.de/Redaktion/DE/Artikel/Energie/sinteg.html
[16] https://www.kopernikus-projekte.de/start
[17] https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/ErneuerbareEnergien/ZahlenDatenInformationen/zahlenunddaten-node.html
[18] https://www.tagesschau.de/inland/cyberangriffe-107.html
[19] BDEW und rheingold institut: Digitalisierung aus Kundensicht, 22. 3. 2017, https://www.bdew.de/digitalisierung-aus-kundensicht

Advanced Software Engineering

21

Developing and testing model-based software securely and efficiently

Prof. Dr. Ina Schieferdecker · Dr. Tom Ritter
Fraunhofer Institute for Open Communication Systems FOKUS

Summary

Software rules them all! In every industry now, software plays a dominant role in technical and business innovations, in improving functional safety, and also for increasing convenience. Nevertheless, software is not always designed, (re) developed, and/or secured with the necessary professionalism, and there are unnecessary interruptions in the development, maintenance, and operating chains that adversely affect reliable, secure, powerful, and trustworthy systems. Current surveys such as the annual World Quality Report put it bluntly, directly correlated with the now well-known failures of large-scale, important and/or safety-critical infrastructures caused by software. It is thus high time that software development be left to the experts and that space be created for the use of current methods and technologies. The present article sheds light on current and future software engineering approaches that can also and especially be found in the Fraunhofer portfolio.

21.1 Introduction

Let us start with the technological changes brought about by the digital transformation which, in the eyes of many people, represent revolutionary innovations for our society. Buildings, cars, trains, factories, and most of the objects in our everyday lives are either already, or will soon be, connected with our future gigabit society via the ubiquitous availability of the digital infrastructure [1]. This will change information sharing, communication, and interaction in every field of life and work,

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_21

be it in healthcare, transport, trade, or manufacturing. There are several terms used to describe this convergence of technologies and domains driven by digital networking: the Internet of Things, smart cities, smart grid, smart production, Industry 4.0, smart buildings, the Internet of systems engineering, cyber-physical systems, or the Internet of Everything. Notwithstanding the different aims and areas of application, the fundamental concept behind all of these terms is the all-encompassing information sharing between technical systems – digital networking:

Digital networking is the term used to refer to the continuous and consistent linking of the physical and digital world. This includes digital recording, reproduction, and modelling of the physical world as well as the networking of the resulting information. This enables real-time and semi-automated observation, analysis, and control of the physical world.

Digital networking facilitates seamless sharing of information between the digital representations of people, things, systems, processes, and organizations and develops a global network of networks – an *inter-net* – that goes far beyond the vision of the original Internet. But this new form of network is no longer a matter of networking for its own sake. Instead, individual data points are combined into information in order to develop globally networked and networkable knowledge and utilize this both for increasing understanding as well as for the management of monotonous or safety critical processes.

In light of this digital networking, the central role of software continues to increase. Digital reproductions – the structures, data, and behavioral models of things, systems, and processes in the physical world – are all realized via software. But so are also all of the algorithms with which these digital reproductions are visualized, interpreted, and reprocessed, as well as all of the functions and services of the infrastructures and systems such as servers and (end) devices in the network of networks. Until recently the essential characteristics of the infrastructures and systems were defined by the characteristics of the hardware, and it was largely a matter of software and hardware co-design. Now the hardware is moving into the background due to generic hardware platforms and components and is being defined by software or even virtualized from the user's point of view. Current technical developments here are software defined networks including network slices, or cloud services such as Infrastructure as a Service, Platform as a Service, or Software as a Service.

In addition, these software-based systems today significantly influence critical infrastructures such as electricity, water, or emergency care: they are an integral part of the systems such that both the software contained or used as well as the infrastructures themselves become so-called critical infrastructure. Here, we are using the term "software-based system" as an overarching term for the kinds of systems whose functionality, performance, security, and quality is largely defined by software. These include networked and non-networked control systems such as control units in automobiles and airplanes, systems for connected and autonomous driving, and systems of systems such as the extension of the automobile into the backbone infrastructure of the OEMs. But also systems (of systems) in telecommunications networks, IT, industrial automation, and medical technologies are understood by this term.

Software-based systems today are often distributed and connected, are subject to real-time demands (soft or hard), are openly integrated into the environment via their interfaces, interact with other software-based systems, and use learning or autonomous functionalities to master complexity. Independently of whether we are now in a fourth revolution or in the second wave of the third revolution with digitization, the ongoing convergence of technologies and the integration of systems and processes is brought about and supported via software. New developments such as those in augmented reality, fabbing, robotics, data analysis, and artificial intelligence, too, place increasing demands on the reliability and security of software-based systems.

21.2 Software and software engineering

Let us examine things in greater depth. According to the IEEE Standard for Configuration Management in Systems and Software Engineering (IEEE 828-2012 [1]), software is defined as “computer programs, procedures and possibly associated documentation and data pertaining to the operation of a computer system”. It includes programmed algorithms, data capturing or representing status and/or context, and a wide range of descriptive, explanatory, and also specifying documents (see Fig. 21.1).

A look at current market indicators reveals the omnipresence of software: according to a 2016 Gartner study, global IT expenditures of $3.5 billion were expected in 2017. Software is thus the fastest-growing area, at $357 billion or 6% [4]. Bitkom, as well, supports this view [5]: according to its own survey of 503 companies with 20 or more staff, every third company in Germany is developing its own software. Among large organizations with 500 or more staff, the proportion rises as high as to 64%. According to this survey, already every fourth company in Germany employs software developers, and an additional 15% say they want to hire additional software specialists for digital transformation.

Nevertheless, 50 years after the software crisis was explicitly addressed in 1968, and after numerous approaches and new methods in both software and quality engineering, the development and operation of software-based connected systems is still not smooth [8]. The term “software engineering” was initially introduced by

F. L. Bauer as a provocation: "the whole trouble comes from the fact that there is so much tinkering with software. It is not made in a clean fabrication process, which it should be. What we need, is software engineering." The authors Fitzgerald and Stol identify various gaps in the development, maintenance, and distribution of software-based systems that can be closed via methods of continuous development, testing, and rollout.

Studies on breakdowns and challenges in the Internet of Things (IoT) complete our view here: according to self-reports by German companies, four in five of them have an "availability and data security gap" in IT services [9]. Servers in Germany, for example, stand idle for an average of 45 minutes during an outage. The estimated direct costs of these kinds of IT failures rose by 26% in 2016 to $21.8 million, versus $16 million in 2015. And these figures do not include the impacts that cannot be precisely quantified such as reduced customer confidence or reputational damage to the brand.

The top two challenges connected to IoT are security in particular IT security and data protection, as well as functional safety and interoperability of the software-defined protocol and service stacks [10].

In keeping with this, the latest 2016–17 edition of the World Quality Report [3] shows that there is a change in the primary goals of those responsible for quality assurance and testing that is accompanying the ongoing pervasion of the physical world by the digital world with the Internet of Things. The change picks up the increasing risk of breakdown and the criticality of software-based connected systems from the perspective of business and security. Thus, increasing attention is given to quality and security by design, and the value of quality assurance and testing is being raised in spite of, or indeed due to, the increasing utilization of agile and DevOps methods. Thus, with the complexity of software-based connected systems, expenditures for the realization, utilization, and management of (increasingly virtualized) test environments are also increasing. Even though extensive cost savings are equally possible in this area through automation, the necessity of making quality assurance and testing even more effective at all levels remains.

21.3 Selected characteristics of software

Before turning to current approaches to developing software-based connected systems, let us first take a look at the characteristics of software. Software should be understood as a technical product that must be systematically developed using software engineering. Software is characterized by its functionality and additional qualitative features such as reliability, usability, efficiency, serviceability, compatibility, and portability [12]. Against the backdrop of current developments and revelations, aspects of ethics as well as of guarantees and liability must also supplement the dimensions of software quality.

For a long time, software was considered to be free from all of the imponderables inherent to other technical products, and in this way was seen as the ideal technical product [11]. A key backdrop to this is the fact that algorithms, programming concepts and languages, and thus any computability is traced to the Turing computability (the Turing thesis). According to Church's thesis, computability here incorporates precisely those functions which can be calculated intuitively by us. Thus, while non-computable problems such as the halting problem elude algorithmics (and thus software), for each intuitively computable function there is an algorithm with limited computing complexity that can be realized via software. Here, the balance between function, algorithm, and software is the responsibility of various phases and methods of software engineering such as specification, design, and implementation as well as verification and validation. If alongside this understanding of intuitive computability, software now sounds like a product that is simple to produce, this is by no means the case. What began with the software crisis still holds true today. Herbert Weber reiterated this in 1992, "the so-called software crisis has thus far not yet produced the necessary level of suffering required to overcome it" [13]. Also Jochen Ludewig in 2013 formulated it as, "the requirements of software engineering have thus not yet been met" [11]. The particular characteristics of software are also part of the reason for this.

First and foremost, software is immaterial, such that all of the practical values for material products do not apply or are only transferable in a limited sense. Thus, software is not manufactured but "only" developed. Software can be copied practically without cost, with the original and the copy being completely identical and impossible to distinguish. This leads, among other things, to nearly unlimited possibilities for reusing software in new and typically unforeseen contexts.

On the one hand, using software does not wear it out. On the other hand, the utilization context and execution environment of software are constantly evolving such that untarnished software does in fact age technologically and indeed logically and thus must be continually redeveloped and modernized. This leads to maintenance cycles for software that, instead of restoring the product to its original state, generate new and sometimes ill-fitting, i.e. erroneous, conditions.

Software errors thus do not arise from wear and tear to the product but are built into it. Or errors develop in tandem with the software's unplanned use outside of its technical boundary conditions. This is one way that secure software can be operated insecurely, for example.

In addition, the days of rather manageable software in closed, static, and local use contexts for mainly input and output functionalities are long gone. Software is largely understood as a system built on distributed components with open interfaces. The components of these can be realized in various technologies and by various manufacturers, and with configurations and contexts that may change dynamically. These may further incorporate third-party systems flexibly by means of service orchestrations and various kinds of interface and network access, which must be able to serve various usage scenarios and load profiles. Actions and reactions cannot be described by consistent functions.

Our understanding of intuitive computability is being challenged daily by new concepts such as data-driven, autonomous, self-learning, or self-repairing software. In doing so, software is increasingly using heuristics for its decision-making in order to efficiently arrive at practicable solutions, even in the case of NP-complete problems. The bottom line is that software-based connected systems, with all of the elementary decision-making they incorporate, are highly complex – the most complex technical systems that have yet been created. In this process potential difficulties arise, simply due to the sheer size of software packages. Current assessments of selected open-source software packages, for example, reveal relationships between software complexity, "code smells", which are indicators of potential software defects, and software vulnerabilities, the software's weak points with respect to IT security. This relationship may not be directly causal but is nevertheless identifiable and worthy of further investigation [14].

21.4 Model-based methods and tools

In what follows, we illustrate selected model-based methods and tools for the efficient development of reliable and secure software that are the result of current R&D studies at Fraunhofer FOKUS.

Models have a long tradition in software development. They originally served the specification of software and its formal verification of correctness. In the meantime they are commonly used as abstract, technology-independent bearers of information for all aspects of software development and quality assurance [15]. They thus serve to mediate information between software tools and to provide abstractions for capturing complex relationships. One example of this, in the context of risk analysis and assessment, or of systematic measurement and visualization of software characteristics, and also of software test automation, is via model-based testing or test automation. As Whittle, Hutchinson, and Rouncefield argue in [16], the particular added value of model-driven software development (Model-Driven Engineering, MDE) is the specification of the architectures, interfaces, and components of software. Architecture is also used by FOKUS as the foundation for documentation, functionality, interoperability, and security in the methods and tools introduced in what follows.

Process automation

Modern software development processes often use teams at various sites for individual components and to integrate commercial third-party components or open source software. Here, a wide range of software tools are used, with various individuals participating in different roles, whether actively or passively. The central problems here are the lack of consistency of the artifacts created in the development process, the shortage of automation, and the lack of interoperability between the tools.

ModelBus® is an open-source framework for tool integration in software and system development and closes the gap between proprietary data formats and software tool programming interfaces [17]. It automates the execution of tedious and error-prone development and quality assurance tasks such as consistency assurance across the entire development process. To do this, the framework uses service orchestrations of tools in keeping with SOA (service-oriented architecture) and ESB (enterprise service bus) principles.

The software tools of a process landscape are connected to the bus by the provision of ModelBus® adaptors. Adaptors are available for connecting IBM Rational Doors, Eclipse and Papyrus, Sparx Enterprise Architect, Microsoft Office, IBM Rational Software Architect, IBM Rational Rhapsody, or MathWorks Matlab Simulink.

21.5 Risk assessment and automated security tests

Safety-critical software-based systems are subject to careful risk analysis and evaluation according to the ISO Risk Management Standard [18] in order to capture and minimize risks. For complex systems, however, this risk management may be very time-consuming and difficult. While the subjective assessment of experienced experts may be an acceptable method of risk analysis on a small scale, with increasing size and complexity other approaches such as risk-based testing [21] need to be chosen.

An additional opportunity for more objective analysis is provided by the use of security tests in line with ISO/IEC/IEEE “Software and systems engineering – Software testing” (ISO 29119-1, [19]). A further option is to first have experts carry out a high-level assessment of the risks based on experience and literature. In order to make this initial risk assessment more accurate, security tests can be employed at precisely the point where the first high-level risk picture shows the greatest vulnerabilities. The objective test results may then be used to enhance, refine, or correct the risk picture thus far. However, this method first becomes economically applicable with appropriate tool support.

RACOMAT is a risk-management tool developed by Fraunhofer FOKUS, which in particular combines risk assessment with security tests [20]. Here, security testing can be directly incorporated into event simulations that RACOMAT uses to calculate risks. RACOMAT facilitates extensive automation of risk modelling through to security testing. Existing databases such as those of known threat scenarios are used by RACOMAT to ensure a high degree of reuse and avoid errors.

At the same time, RACOMAT supports component-based compositional risk assessment. Easy-to-understand risk graphs are used to model and visualize an image of the risk situation. Common techniques such as fault tree analysis (FTA), event tree analysis (ETA), and conducting security risk analysis (CORAS) may be used in combination for the risk analysis in order to be able to benefit from the various strengths of the individual techniques. Starting with an overall budget for risk assessment, RACOMAT calculates how much expenditure is reasonable for security testing in order to improve the quality of the risk picture by reducing vul-

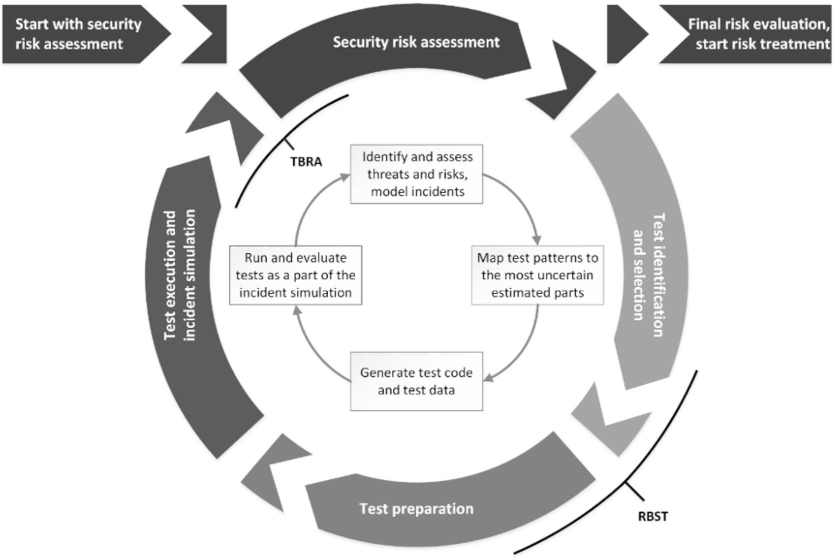


Fig. 21.1 Risk analysis and security testing with RACOMAT (Fraunhofer FOKUS)

nerabilities. The tool offers recommendations on how these means should be used. To do this, RACOMAT identifies relevant tests and places them in order of priority.

21.6 Software mapping and visualization

Software-based systems are becoming ever more complex due to their increasing functions and their high security, availability, stability, and usability requirements. In order for this not to lead to losses of quality and so that structural problems can be identified early on, quality assurance must commence right at the beginning of the development process. A model-driven development process where models are key to the quality of the software-based system is well suited to this. Up to now, however, quality criteria for this were neither defined nor established. In future, model characteristics and their quality requirements need to be identified, and additionally mechanisms found with which their properties and quality can be determined.

Metrino is a tool that checks and ensures the quality of models [22]. It may be used in combination with ModelBus® but can also be employed independently. With the aid of Metrino, metrics for domain-specific models can be generated, independently defined, and managed. The metrics produced can be used for all models that accord with the meta-model used as the basis for development. Metrino thus analyzes and verifies properties such as the complexity, size, and description of software artifacts. In addition, the tool offers various possibilities for checking the computational results of the metrics and representing them graphically – for ex-

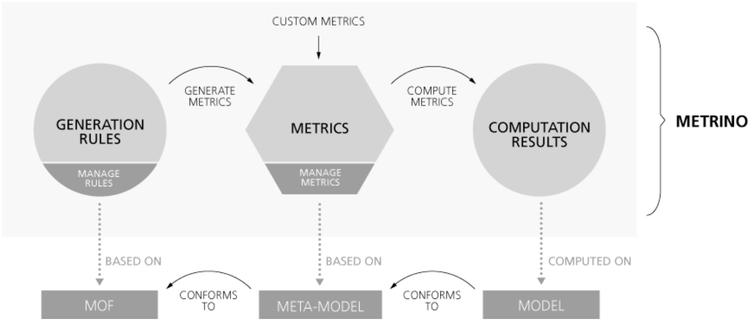


Fig. 21.2 Model-based software measurement and visualization with Metrino (Fraunhofer FOKUS)

ample in a table or spider chart. Since Metrino saves the results from several evaluations, results from different time periods can also be analyzed and compared with one another. This is the only way that optimal quality of the final complex software-based system can be guaranteed.

Metrino is based on the Structured Metrics Metamodel (SMM) developed by the Object Management Group (OMG) and can be used both for models in the Unified Modeling Language (UML) as well as for domain-specific modelling languages (DSLs). On top of that, Metrino's field of application includes specialized, tool-specific languages and dialects.

Whether for designing embedded systems or for software in general, Metrino can be used in the widest variety of different domains. The tool can manage metrics and apply them equally to (model) artifacts or also to the complete development chain, including traceability and test coverage.

21.7 Model-based testing

The quality of products is the decisive factor of being accepted on the market. In markets with security-related products in particular, such as medicine, transportation, or automation sectors, for example, quality assurance is thus accorded high priority. In these sectors, quality is equally decisive for product authorization. Quality assurance budgets, however, are limited. It is thus important to managers and engineers that the available resources are utilized efficiently. Often, manual testing methods are still being used, even if only a comparatively small number of tests can be prepared and conducted in this way, and they are additionally highly prone to error. The efficiency of manual testing methods is thus limited, and rising costs are unavoidable. Model-based test generation and execution offers a valuable alternative: the use of models from which test cases can be automatically derived offers enormous potential for increasing test quality at lower costs. In addition, case studies and practical uses have shown that when model-based testing techniques are introduced necessary investment costs in technology and training pay off quickly [24].

Fokus!MBT thus offers an integrated test modelling environment that leads the user through the Fokus!MBT methodology and thus simplifies the creation and use of the underlying test model [25]. A test model contains test-relevant, structural, behavior- and method-specific information that conserves the tester's knowledge in a machine processable fashion. In this way, it can be adapted or evaluated at any time, say for the generation of additional test-specific artifacts. Additional benefits of the test model are the visualization and documentation of the test specification.

Fokus!MBT uses the UML Testing Profile (UTP), specified by the Object Management Group and developed with significant contributions from FOKUS, as its modeling notation. UTP is a test-specific extension of the Unified Modeling Language (UML) common in industry. This allows testers to use the same language concepts as the system architects and requirements engineers, thus preventing communication issues and encouraging mutual understanding.

Fokus!MBT is based on the flexible Eclipse RCP platform, the Eclipse Modeling Framework (EMF), and Eclipse Papyrus. As a UTP-based modeling environment, it has all of the UML diagrams available as well as additional test-specific diagrams. Alongside the diagrams, Fokus!MBT relies on a proprietary editor framework for describing and editing the test model. The graphical editor user interfaces can be specifically optimized for the needs or abilities of the user in question. In doing so, if necessary, it is possible to completely abstract from UML/UTP, allowing specialists unfamiliar with IT to quickly produce a model-based test specification. This is also supported by the provision of context-specific actions that lead the user through the Fokus!MBT methodology. In this way, methodically incorrect actions or actions, which are inappropriate for the context in question are not even enabled. Based upon this foundation, Fokus!MBT integrates automated modeling rules that guarantee adherence to guidelines, in particular modelling or naming conventions, both after and during working on the test model. These preventative quality assurance mechanisms

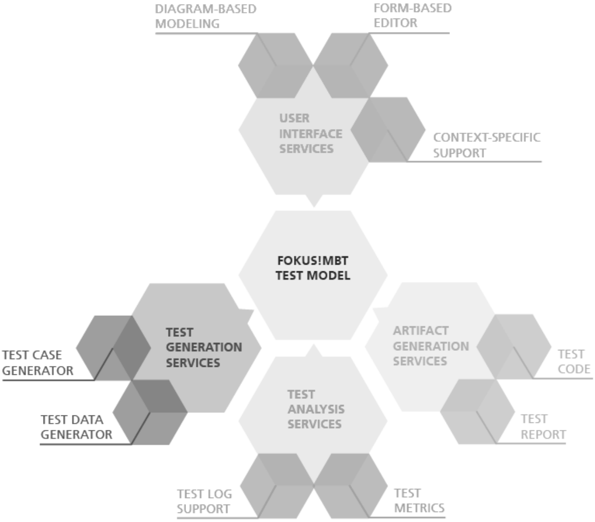


Fig. 21.3 Model-based testing with Fokus!MBT (Fraunhofer FOKUS)

distinguish Fokus!MBT from other UML tools, accelerate model generation, and minimize costly review sessions.

The fundamental goal of all test activities is validating the system to be tested vis-à-vis its requirements. Consistent and uninterrupted traceability, in particular between requirements and test cases, is indispensable here. Fokus!MBT goes one step further and also incorporates the test execution results into the requirements traceability within the test model. In this way, a traceability network is created between requirement, test case, test script, and test execution result, thus making the status of the relevant requirements or the test progress immediately assessable. The visualization of the test execution results additionally facilitates the analysis, processing and assessment of the test case execution process. The test model thus contains all of the relevant information to estimate the quality of the system tested, and support management in their decision-making related to the system's release.

21.8 Test automation

Analytical methods and dynamic testing approaches in particular are a central and often also exclusive instrument for verifying the quality of entire systems. Software tests thereby require all of the typical elements of software development, because tests themselves are software-based systems and must thus be developed, built, tested, validated, and executed in exactly the same way. In addition to that, test systems possess the ability to control, stimulate, observe and to assess the system being tested. Although standard development and programming techniques are generally also applicable for tests, specific solutions for the development of a test system are important in order to be able to take its unique features into account. This approach expedited the development and standardization of specialized test specification and test implementation languages.

One of the original reasons for the development of Tree and Tabular Combined Notation (TTCN) was the precise conformity definition for telecommunications component protocols according to their specification. Test specifications were utilized to define test procedures objectively and assess, compare, and certify the equipment on a regular basis. Thus, the automated execution became exceptionally important for TTCN, too.

Over the years, the significance of TTCN grew and various pilot projects demonstrated a successful applicability beyond telecommunications. With the convergence of telecommunications and information technology sectors, the direct applicability of TTCN became obvious to developers from other sectors. These trends, along with the characteristics of more recent IT and telecommunications technolo-

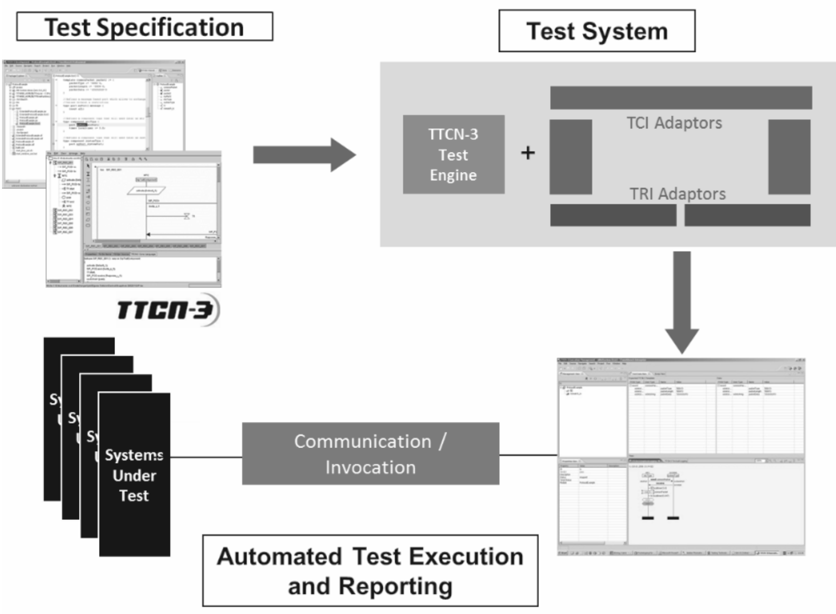


Fig. 21.4 Test automation with TTCN-3 (Fraunhofer FOKUS)

gies also placed new requirements on TTCN: the result is TTCN-3 (Testing and Test Control Notation Version 3, [27]).

TTCN-3 is a standardized and modern test specification and test implementation language developed by the European Telecommunication Standards Institute (ETSI). Fraunhofer FOKUS played a key role in TTCN-3's development and is responsible for various elements of the language definition, including Part 1 (concepts and core languages), Part 5 (runtime interfaces), and Part 6 (test control interfaces), as well as TTCN-3 tools and test solutions [28][29]. With the aid of TTCN-3, tests can be developed textually or graphically, and execution can be automated. In contrast to many (test) modeling languages, TTCN-3 comprises not only a language for test specification but also an architecture and execution interfaces for TTCN-3-based test systems. Currently, FOKUS uses TTCN-3 for developing the Eclipse IoT-testware for testing and securing IoT components and solutions, for example [30].

21.9 Additional approaches

It is not possible to introduce all of our methods, technologies, and tools here. Our publications (see also [31]) contain further information on

- Security-oriented architectures,
- Testing and certifying functional security,
- Model-based re-engineering,
- Model-based documentation,
- Model-based porting to the cloud, or
- Model-based fuzz tests.

21.10 Professional development offerings

It is not enough to simply develop new methods, technologies, and tools. These also need to be distributed and supported during their introduction to projects and pilots.

Fraunhofer FOKUS has thus for a long time been involved in professional development. The institute has initiated and/or played a key role in developing the following professional development schemes in cooperation with the ASQF (Arbeitskreis Software-Qualität und -Fortbildung – "Software Quality and Training Working Group") [32], GTB (German Testing Board) [33], and the ISTQB (International Testing Qualifications Board [34]):

- GTB Certified TTCN-3 Tester
- ISTQB Certified Tester Foundation Level – Model-Based Testing
- ISTQB Certified Tester Advanced Level – Test Automation Engineer
- ASQF/GTB Quality Engineering for the IoT

Further, Fraunhofer FOKUS, together with HTW Berlin and Brandenburg University of Applied Sciences, is also forming a consortium, the Cybersecurity Training Lab [35], with training modules on

- Secure software engineering
- Security testing
- Quality management & product certification
- Secure e-government solutions
- Secure public safety solutions

This and other offerings, such as on semantic business rule engines or open government data, are also available via the FOKUS-Akademie [36].

21.11 Outlook

Software development and quality assurance are both subject to the competing requirements of increasing complexity, the demand for high-quality, secure, and reliable software, and the simultaneous economic pressure for short development cycles and fast product introduction times.

Model-based methods, technologies, and tools address the resulting challenges, and in particular support modern agile development and validation approaches. Continuous development, integration, testing and commissioning benefit from model-based approaches to a particular degree. This is because they form a strong foundation for automation and can also support future technology developments due to their independence from specific software technologies.

Additional progress in model-based development is to be expected or indeed forms part of current research. Whereas actual integration and test execution are already conducted in a nearly entirely automated fashion, the analysis and correction of errors remains a largely manual task, one that is time-consuming and itself subject to error and can thus lead to immense delays and costs. Self-repairing software would be an additional step towards greater automation, borrowing from the diverse software components in open source software using pattern recognition and analysis through deep learning methods, and repair and assessment using evolutionary software engineering approaches. In this way, software could become not only self-learning but also self-repairing.

Nevertheless, until then it is important to

- Understand software engineering as an engineering discipline and leave it to experts to develop and to ensure their quality including safety and/or security,
- Continue to develop software engineering itself as a field of research and development and automate insecure manual steps in software development and validation,
- Consider beginning with draft monitoring and testing environments for all levels of a digitized application landscape that can be efficiently managed via virtualization methods for software platforms,
- Consider that security, interoperability, and usability are gaining increasingly in importance in the quality of software-based connected systems and demand priority during design, development, and validation.

Sources and literature

[1] Henrik Czernomoriez, et al.: Netzinfrastrukturen für die Gigabitgesellschaft, Fraunhofer FOKUS, 2016.

[2] IEEE: IEEE Standard for Configuration Management in Systems and Software Engineering, IEEE, 828-2012, https://standards.ieee.org/findstds/standard/828-2012.html, besucht am 15.7.2017.

[3] World Quality Report 2016-2017: 8th Edition – Digital Transformation, http://www.worldqualityreport.com, besucht am 15.7.2017.

[4] Gartner 2016: Gartner Says Global IT Spending to Reach $3.5 Trillion in 2017, http://www.gartner.com/newsroom/id/3482917, besucht am 15.7.2017.

[5] Bitkom Research 2017: Jedes dritte Unternehmen entwickelt eigene Software, https://www.bitkom.org/Presse/Presseinformation/Jetzt-wird-Fernsehen-richtig-teuer.html, besucht am 15.7.2017.

[6] NATO Software Engineering Conference, 1968: http://homepages.cs.ncl.ac.uk/brian.randell/NATO/nato1968.PDF, besucht am 21.7.2017.

[7] Friedrich L. Bauer: Software Engineering – wie es begann. Informatik Spektrum, 1993, 16, 259-260.

[8] Brian Fitzgerald, Klaas-Jan Stol, Continuous software engineering: A roadmap and agenda, Journal of Systems and Software, Volume 123, 2017, Pages 176-189, ISSN 0164-1212, http://dx.doi.org/10.1016/j.jss.2015.06.063, besucht am 21.7.2017.

[9] VEEAM: 2017 Availability report, https://go.veeam.com/2017-availability-report-de, besucht am 21.7.2017.

[10] Eclipse: IoT Developer Trends 2017 Edition, https://ianskerrett.wordpress.com/2017/04/19/iot-developer-trends-2017-edition/, besucht am 21.7.2017.

[11] Jochen Ludewig und Horst Lichter: Software Engineering. Grundlagen, Menschen, Prozesse, Techniken. 3., korrigierte Auflage, April 2013, dpunkt.verlag, ISBN: 978-3-86490-092-1.

[12] ISO/IEC: Systems and software engineering -Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models, ISO/IEC 25010:2011, https://www.iso.org/standard/35733.html, besucht am 22.7.2017.

[13] Herbert Weber: Die Software-Krise und ihre Macher, 1. Auflage, 1992, Springer-Verlag Berlin Heidelberg, DOI 10.1007/978-3-642-95676-8.

[14] Barry Boehm, Xavier Franch, Sunita Chulani und Pooyan Behnamghader: Conflicts and Synergies Among Security, Reliability, and Other Quality Requirements. QRS () 2017 Panel, http://bitly.com/qrs_panel, besucht am 22.7.2017.

[15] Aitor Aldazabal, et al. „Automated model driven development processes." Proceedings of the ECMDA workshop on Model Driven Tool and Process Integration. 2008.

[16] Jon Whittle, John Hutchinson, and Mark Rouncefield. „The state of practice in model-driven engineering." IEEE software 31.3 (2014): 79-85.

[17] Christian Hein, Tom Ritter und Michael Wagner: Model-Driven Tool Integration with ModelBus. In Proceedings of the 1st International Workshop on Future Trends of Model-Driven Development – Volume 1: FTMDD, 35-39, 2009, Milan, Italy.

[18] ISO: Risk management, ISO 31000-2009, https://www.iso.org/iso-31000-risk-management.html, besucht am 22.7.2017.

[19] ISO/IEC/IEEE: Software and systems engineering – Software testing – Part 1: Concepts and definitions. ISO/IEC/IEEE 29119-1:2013, https://www.iso.org/standard/45142.html, besucht am 22.7.2017.
[20] Johannes Viehmann und Frank Werner. „Risk assessment and security testing of large scale networked systems with RACOMAT.“ International Workshop on Risk Assessment and Risk-driven Testing. Springer, 2015.
[21] Michael Felderer, Marc-Florian Wendland und Ina Schieferdecker. „Risk-based testing.“ International Symposium On Leveraging Applications of Formal Methods, Verification and Validation. Springer, Berlin, Heidelberg, 2014.
[22] Christian Hein, et al. „Generation of formal model metrics for MOF-based domain specific languages.“ Electronic Communications of the EASST 24(2010).
[23] Marc-Florian Wendland, et al. „Model-based testing in legacy software modernization: An experience report.“ Proceedings of the 2013 International Workshop on Joining AcadeMiA and Industry Contributions to testing Automation. ACM, 2013.
[24] Ina Schieferdecker. „Model-based testing.“ IEEE software 29.1 (2012): 14.
[25] Marc-Florian Wendland, Andreas Hoffmann, and Ina Schieferdecker. 2013. Fokus!MBT: a multi-paradigmatic test modeling environment. In Proceedings of the workshop on ACadeMics Tooling with Eclipse (ACME ,13), Davide Di Ruscio, Dimitris S. Kolovos, Louis Rose, and Samir Al-Hilank (Eds.). ACM, New York, NY, USA, Article 3, 10 pages. DOI: https://doi.org/10.1145/2491279.2491282
[26] ETSI: TTCN-3 – Testing and Test Control Notation, Standard Series ES 201 873-1 ff.
[27] Jens Grabowski, et al. „An introduction to the testing and test control notation (TTCN-3).“ Computer Networks 42.3 (2003): 375-403.
[28] Ina Schieferdecker und Theofanis Vassiliou-Gioles. „Realizing distributed TTCN-3 test systems with TCI.“ Testing of Communicating Systems (2003): 609-609.
[29] Juergen Grossmann, Diana Serbanescu und Ina Schieferdecker. „Testing embedded real time systems with TTCN-3.“ Software Testing Verification and Validation, 2009. ICST‘09. International Conference on. IEEE, 2009.
[30] Ina Schieferdecker, et al. IoT-Testware – an Eclipse Project, Keynote, Proc. of the 2017 IEEE International Conference on Software Quality, Reliability & Security, 2017.
[31] FOKUS: System Quality Center, https://www.fokus.fraunhofer.de/sqc, besucht am 22.7.2017.
[32] ASQF: Arbeitskreis Software-Qualität und Fortbildung (ASQF), http://www.asqf.de/, besucht am 25.7.2017.

Automated Driving

22

Computers take the wheel

Prof. Dr. Uwe Clausen
Fraunhofer Institute for Material Flow and Logistics IML
Prof. Dr. Matthias Klingner
Fraunhofer Institute for Transportation and Infrastructure Systems IVI

Summary

Digital networking and autonomous driving functions mark a new and fascinating chapter in the success history of automobile manufacture, which stretches back well over a century already. With powerful environment recognition, highly accurate localization and low-latency communication technology, vehicle and traffic safety will thus increase dramatically. Precise, fully automated vehicle positioning of autonomously driven electric cars creates the conditions for introducing innovative high-current charging technologies located underground. If autonomous or highly automated vehicles share information with intelligent traffic controls in future, then this may lead to a significantly more efficient utilization of existing traffic infrastructures and marked reductions in traffic-related pollutant emissions. These are three examples that underscore the enormous significance of electric mobility together with autonomous driving functions for the development of a truly sustainable mobility. In this process, the range of scientific-technical challenges in need of a solution is extraordinarily broad. Numerous Fraunhofer institutes are involved in this key development process for our national economy, contributing not merely expert competencies at the highest scientific-technical level, but also practical experience in the industrial implementation of high technologies. In what follows, we take a look at some of the current topics of research. These include autonomous driving functions in complex traffic situations, cooperative driving maneuvers in vehicle swarms, low-latency communication, digital maps and precise localization. Also, security of functions and security against manipulation for driverless vehicles, digital

Reimund Neugebauer, *Digital Transformation*
https://doi.org/10.1007/978-3-662-58134-6_22

networking and data sovereignty in intelligent traffic systems is considered. Finally, range extension and fast-charging capabilities for autonomous electric vehicles through to new vehicle design, modular vehicle construction and scalable functionality is addressed. And even though the automobile sector is the focus of our attention here, it is worth taking a look at interesting Fraunhofer developments in autonomous logistics transport systems, driverless mobile machines in agricultural engineering, autonomous rail vehicle technology, and unmanned ships and underwater vehicles.

22.1 Introduction

The initial foundations for highly automated driving functions were already developed more than 20 years ago under the European PROMETHEUS project (PROgraMme for a European Traffic of Highest Efficiency and Unprecedented Safety, 1986– 1994). With €700 m. in funding, it was Europe's largest ever research project to date and involved not only the vehicle manufacturers but also nearly all of the main European supply firms and academic institutions. The goal was to improve the efficiency, environmental sustainability, and security of road traffic. Fraunhofer institutes such as the then IITB and today's IOSB in Karlsruhe have been able to successfully continue this research right up to the present day. Many of the findings from these first years have now found broad application in modern vehicle technology. These include, for example, proper vehicle operation, friction monitoring and vehicle dynamics, lane keeping support, visibility range monitoring, driver status monitoring or system developments in collision avoidance, co-operative driving, autonomous intelligent cruise control, and automatic emergency calls.

Then as now, the migration to ever higher degrees of automation is closely linked to the development of driver assistance systems. Emergency braking assistants, multi collision brakes, pre-crash security systems, and lane-keeping and lane-change assistants all offer ever more comprehensive automated driving functions for vehicle occupant protection. While these driver assistance systems primarily actively intervene in vehicle control in security-critical driving situations, autonomous cruise control (ACC) relieves the driver in monotonous traffic flows on expressways and highways. Based on ACC and lane-keeping systems together with a high-availability car-to-car communication, the first forms of platooning are currently developing. Here, vehicles are guided semi-autonomously as a group over longer expressway distances at high speed and with minimal safety distances between them. It is commonly expected that within the coming years platooning will continue to be developed for utility vehicles in particular and be transitioned into practical use.

22.2 Autonomous driving in the automobile sector

22.2.1 State of the art

In the automobile sector, too, automated expressway driving and automatic maneuvering and parking represent challenges that have been largely overcome today. With the addition of active assistive functions for longitudinal and lateral guidance – for example, in the current Mercedes E-Class – vehicles are already shifting into the gray area between semi- and high automation and are thus defining the state of the art for market-ready and legally compliant standard features.

The function-specific form of the most highly developed approaches for automated driving functions was first demonstrated during the DARPA Grand Challenge hosted by the Defense Advanced Research Projects Agency (DARPA) of the US Defense Department. This was a competition for unmanned land vehicles, last held in 2007, that significantly advanced the development of vehicles capable of driving completely autonomously.

Since then, the IT corporation Google (now Alphabet) has become one of the technological leaders in the field of autonomous vehicles. As of November 2015, Google's driverless cars reached the mark of 3.5 million test kilometers with a new record proportion of 75% of these being fully autonomous. Google currently operates around 53 vehicles capable of driving completely automatically using a combination of (D)GPS, laser scanners, cameras, and highly accurate maps. Even if the safety of the autonomous driving functions implemented still requires a certain amount of verification, Tesla – with sales of 500,000 highly automated passenger vehicles per year – is also planning to set new standards in the field of highly automated electric vehicles from 2018. Established vehicle manufacturers such as BMW, Audi, and Nissan have thus far primarily demonstrated autonomous driving maneuvers within limited areas such as expressways and parking garages. In 2013, Mercedes-Benz showed what can be implemented today under ideal conditions, with close-to-production sensors and actuators, with its legendary autonomous cross-country trip from Mannheim to Pforzheim. Swedish vehicle manufacturer Volvo, itself active in the field of autonomous driving, is currently implementing a cloud-based road information system in Sweden and Norway. The rapid sharing of highly accurate digital map content, road status information, and current traffic data is a key requirement for ensuring adequate safety for future autonomous driving.

Alongside automobile manufacturers, large tier-one suppliers such as ZF and Continental are also active in the field of automation and are increasingly beginning to present their own solutions. The first markets for prototype driverless shuttle vehicles are already beginning to open up.

The demonstrations by OEMs, suppliers, and shuttle manufacturers described above show that completely automatic driving will be possible even with production vehicles in the medium term. Here, less complex environments such as expressways are the initial focus, especially when considering production vehicle development, and the safety driver serves as the supervisory entity. This concept is only suitable to a limited extent for comprehensively guaranteeing the safety of highly automated vehicles. The fact that the spontaneous transfer of the driving function to the driver in genuinely critical situations can lead to driving mistakes has been underscored by the first, partly tragic accidents. Centrally operated roadside vehicle monitoring with externally synchronized vehicle movement in situations of extreme danger represents one alternative to the safety driver, but has thus far received little to no consideration from the OEMs.

The Google and Daimler demonstrations show that vehicles can be safely localized with decimeter accuracy in urban environments, using a fusion of odometry, laser sensors, and DGPS. The development of roadside safety systems is thus entirely possible. One such roadside monitoring system – based on lidar systems – is currently being implemented at the University of Reno's Living Lab in the state of Nevada. The Fraunhofer IVI is involved there with corresponding projects. In addition, current developments make clear that the necessary products for networking completely autonomous vehicles with the infrastructure will be widely available in the short to medium term, and thus that the use of these communication structures may also serve to ensure traffic safety.

The growing demand for automated driving functions has been emphasized recently by a number of user studies, conducted by, among others, the Fraunhofer IAO. Surveys by Bitkom, Bosch, Continental, and GfK reveal a high level of interest (33% to 56%) and stress the future significance of the factors of safety and convenience as the most important criteria for new purchases. By 2035, market penetration for automated vehicles of between 20% [1] and 40% [2] is considered a possibility. Just during these next few years, highly developed driver safety systems and convenience-oriented assistance systems will dominate the rapidly growing future car IT market. In their *Connected Car 2014* study, Pricewaterhouse-Coopers estimates an annual growth in global revenues in the two categories mentioned above of around €80 billion by 2020 [3]. The proportion of value added by electronics in cars currently lies at between 35% and 40% of the sale price and rising [4].

The spending of German automotive firms for patents and licenses in the electrics/electronics fields has risen significantly since 2005, currently lying above €2 billion [5]. For in-vehicle connectivity services, McKinsey predicts a fivefold increase in potential sales volumes by 2020 [6]. The combination of these factors makes the automobile market increasingly attractive for companies from the IT sector. In an American Internet study, for example, Google and Intel are already highlighted as the most influential players in the automated driving field [7]. While US firms dominate the IT field, a leading position in the development and introduction of highly advanced automation systems is accorded to German vehicle manufacturers and suppliers. The *Connected Car 2014* study mentioned above identifies three German groups (VW, Daimler, and BMW) as the leading automobile manufacturers for innovation in driver assistance systems. American studies, too, reach similar conclusions, both with respect to Audi, BMW, and Daimler [7], the three premium German manufacturers, as well as the German suppliers Bosch and Continental [3].

22.2.2 Autonomous driving in complex traffic situations

Autonomous driving at VDA standard level 5 (fully driverless) will only be approvable in the public transport space when it is supervised from the roadside by powerful car-to-infrastructure and car-to-car networking, and can be influenced by external safety systems in hazardous situations. Supervised and cooperative driving in so-called automation zones will make its mark in the urban space. This will first take place in the field of inner-city logistics and local public transportation in the form of autonomous shuttle lines, e-taxis, and delivery services. In the medium term, highly automated or fully autonomous vehicles in cooperatively functioning vehicle convoys will enter the stage. In the context of comprehensive automation of inner-city transport networks, these will make a significant contribution to the easing of traffic flows as well as to the more efficient capacity utilization of existing transportation infrastructures.

Currently, various technologies for cooperative driving in automation zones are being developed within Fraunhofer institutes, in particular, institutes in the Fraunhofer ICT Group. Alongside the linear guidance of vehicles following one another (platooning), cooperative driving also includes the multi-lane guidance of heterogeneous swarms of vehicles. Technologies under development are:

- High-performance, reliable communications equipment for autonomous driving functions based on WLAN 11p systems, an extended LTE standard, and 5G mobile telephony
- Monitoring and predictive functions for communications quality
- Fast ad hoc network and protocol technologies
- Robust ad hoc identification and situation detection
- Combined environment sensors based on cameras, radar, lidar, and ultrasound
- Robust and secure sensor data fusion at the perception level
- Interior sensors and structurally integrated electric/electronic systems for the fail-safe capture and forwarding of data and energy allocations
- Fast SLAM technologies for dynamic location correction
- Detection techniques and classifiers for autonomous vehicle guidance
- Machine learning for autonomous driving in the real world
- Cooperative steering and distance regulation
- Robust and secure pathway planning procedures
- Driving strategies and swarm guidance for 2D swarm maneuvers
- Robust methods for roadway condition mapping
- Digital map material with dynamic updates.

As well as established open source robotics frameworks such as ROS, plus YARP, Orocos, CARMEN, Orca, and MOOS, in-house simulation and development tools are also used across institutes here such as OCTANE from Fraunhofer IOSB. Control of the vehicle during autonomous driving is generally transferred to various trajectory controls by implemented finite state machines in keeping with the situation identified. These trajectory controls are responsible for the individual driving tasks such as straight-line driving, lane changes, left- or right-hand turns, etc. Even though impressive outcomes have already been achieved with these situation-dependent trajectory controls, the functional security is not sufficient to model fully automatic driving on public roads. In contrast to the DARPA Challenge's open terrain, real road traffic offers far more complex and sometimes ambiguous situations to overcome, destabilizing the situationally discrete selection processes for allocating the trajectory controls. In [8][9], Fraunhofer IOSB researchers introduced, among other things, a significantly more powerful path planning process that reduces insecurities significantly and is now also being pursued by other authors [10]. For each point in time, the process [10][11] calculates a multicriteria optimized trajectory that incorporates all of the decision-making criteria of a human driver such as collision avoidance, rule conformity, convenience, and journey time.

In probabilistic modeling, not only the fallibility of sensor signals is taken into account but also uncertainties, for example, with respect to the behavior of traffic participants. Using appropriate statistical behavioral models, an autonomous vehicle can thus adapt its driving style to a human's, a key requirement for automated driving in mixed traffic environments. The feasibility of autonomous driving based on this model in mixed traffic environments is also underscored by the results of the EU PROSPECT project. Here, this approach was further developed towards a proactive situational analysis for protecting pedestrians. In [11], Fraunhofer IOSB staff present an optimization process that identifies global optimums for these kinds of driving maneuvers within a specifiable period. [12] showed that these represent a necessary prerequisite for safe autonomous driving in real traffic. During its path planning, an automated vehicle should also take account of the knock-on effects of its maneuvers on the behavior of the vehicles in its environment. This is in particular the case when merging with the flow of traffic, a task that human drivers approach, when they are pulling out of parking spaces, for example, by cautiously creeping forward until one of the vehicles leaves them sufficient space. These kinds of planning processes operate in a very large search space. Highly efficient processes for solving these optimization problems can be found in [13] and [14].

The prerequisite for any kind of autonomous driving is sensor detection of the vehicle environment. In order to meet the high requirements in terms of reliability, field of vision, and range, several different sensor systems are generally used. Using multi-sensor fusion, a consistent environmental model is generated from the measurement data for the purposes of mapping [15][16], obstacle recognition, and for identifying moving objects [17]. The significance of integration technologies for robust, high-resolution environment sensors has already been touched upon. Fraunhofer institutes from the Group for Microelectronics are currently working intensively on new sensors, in particular in the field of suitable radar and lidar systems for automotive use. Without highly integrated sensor technologies based on SiGe and SiGe BiCMOS and the development of enclosure technologies in the millimeter wave range (which Fraunhofer institutes such as EMFT were also involved in developing), large-scale automotive use of radar sensors would not have been possible. In combination with ultrasound, camera, and lidar systems, these radar sensors have now become firmly established. Today's systems largely operate within the globally standardized 76–77 GHz frequency range. 24 GHz sensors are still used for rear applications (blind spot monitoring, lane-change assistants, reversing assistants). Due to bandwidth limitations, a transition to 77 GHz sensors is conceivable here, too.

The diversity of scenarios in the urban environment presents great challenges, first and foremost for video-based environment recognition. Machine learning processes based on deep neural networks (deep learning) have in recent years led to massive improvements. Reductions in the error rate for real-time person detection of up to 80% have been possible, for example [18]. In addition to traditional object detection, it has now been possible to demonstrate processes for fine-grained pixel-level object differentiation [19].

As part of the EU's AdaptIVe research project, 28 different partners – including Europe's ten largest automobile manufacturers, suppliers, research institutes, and universities – are investigating various use cases for autonomous driving on expressways, within car parks, and in complex urban areas. In 2014, the aFAS project commenced, with eight partners from research and industry developing an autonomous security vehicle for expressway roadwork sites. In the IMAGinE project, research is being carried out into the development of new assistance systems for the cooperative behavior context. Within the European SafeAdapt research project, new electric and electronic architecture concepts for increased autonomous electric vehicle reliability are being developed, evaluated, and validated under Fraunhofer's leadership. The C-ITS (Intelligent Transport System) project is drafting a Europe-wide infrastructure and coordinating international cooperation, particularly in the field of communications infrastructures. In 2017, the AUTOPILOT project began, viewing the vehicle as sensor nodes and incorporating it into an IoT structure where the vehicle forms a specific info node.

Tests are ongoing in five different European regions.

22.2.3 Cooperative driving maneuvers

As demonstrated by swarms of fish, birds, and insects and the herd instinct of some mammals, synchronized movements mean that an incredible number of moving objects can be concentrated in an extremely limited space. Well-known examples of spontaneous synchronization outside of the engineering sphere include rhythmic applause or audience waves. Swarm movements in the animal kingdom or the synchronous firing of the synapses in the nervous system; in the engineering sphere, examples include the synchronous running of power generators in a supply system. In the highly automated transport systems of the future, synchronization will

- Increase the traffic density of existing infrastructures,
- Accelerate traffic flows via coordinated prioritization and synchronized signaling systems,
- Extensively optimize transportation chains in public transport via dynamic, synchronized connection conditions and thus
- Significantly reduce traffic-related emissions of CO_2 and pollutants.

Synchronized mobility presupposes a high degree of automation that begins with individual vehicles and extends across the infrastructure to the traffic control centers. The development of synchronization mechanisms for comprehensive traffic management via highly automated dynamic traffic guidance systems includes

- The use of cooperative systems for managing, monitoring, and safeguarding complex traffic flow management systems,
- The implementation of high-dimensional synchronous regulators for dynamically sequencing traffic signals in traffic networks.
- The synchronization of vehicle inflows and outflows in urban areas, and
- The management of synchronously moving vehicle platoons, in-platoon communication, vehicle localization, and distance control.

The Kuramoto model provides a mathematical description of the synchronization of n weakly coupled oscillators with intrinsic natural frequencies ω_i. This non-linear model approach has now found application in a great variety of different scientific fields. It also offers an excellent foundation in transport engineering for forming “dynamic green waves” dependent on the volume of traffic, by introducing a weak, conditional interaction between the individual traffic light circuits. “Weak interactions”, i.e. signals for changing driving profiles or traffic light cycles, are calculated between the vehicles and/or traffic lights depending on local traffic densities and traffic light circuit phasing in real time. This is communicated via radio-based communication channels to the vehicles/traffic lights, and thus utilized for the controlled synchronization of vehicle speeds/traffic light circuits.

The planning of cooperative driving maneuvers is extraordinarily complex, since optimal choices must take account of all of the possible trajectories and combinations for all of the vehicles involved. Algorithms for planning cooperative driving maneuvers, especially for cooperative avoidance of accidents, have been suggested in [20][21][22] and elsewhere. Due to the high degree of problem complexity, various limiting assumptions are applied here, for example, a comparatively coarse discretization of the plans or restriction to specific scenarios [22].

22.2.4 Low-latency broadband communication

Over the course of evolution, the ability of living organisms to respond reflexively to the broadest range of hazardous situations has proved vital to their survival. In the split seconds before an unavoidable collision, pre-crash safety systems (PCSSs) even today protect vehicle occupants and those involved in the accident by tightening the seatbelts, releasing various airbags, or raising the bonnet. Fully autonomous vehicles will in future be able to largely avoid crash situations or, in mixed traffic situations outside of supervised automation zones, at least significantly alleviate them. This is supported by the following:

- Early identification of hazardous situations,
- Car-to-car communication and coordinated trajectory selection,
- Extremely maneuverable chassis design, and
- Highly dynamic and autonomously controlled driving maneuvers.

The safety philosophy governing the development of autonomous vehicles is the assumption that it will remain impossible even in future to identify every hazardous situation sufficiently early, at least for standard evasive or braking maneuvers. For these kinds of dangerous moments, distributed electric drivetrains, multi-axle steering control processes, ABS, ESP and torque vectoring, as well as familiar vehicle stabilization techniques using combined longitudinal/latitudinal control offer additional degrees of freedom. In extreme driving situations, autonomously controlled driving maneuvers will mean that even residual risks can be safely overcome.

Coordinating autonomous driving functions in hazardous situations requires particularly powerful car-to-x communications (LTE-V, 5G) with very short latencies and maximum transmission rates. Fraunhofer institutes such as IIS and HHI are significantly involved in developing and defining standards for these communications technologies.

Which car-to-x communications technologies will eventually prevail over the coming years cannot yet be predicted with certainty. It is possible to see potential development both in technologies already introduced, such as WLAN 11p, as well as in the enhanced 4G standard LTE-V (Vehicular) and LTE Advanced Pro with basic functions from 3GPP Release 14 through to the emerging 5G mobile telephony standard. For narrowband applications, an IoT service was introduced by the Deutsche Telekom in 2017 that facilitates high signal strengths and ranges based on the LTE-Cat-NB1 specification. Fraunhofer IML is involved in the development related to innovative applications in logistics and transportation.

Currently, the communications technology equipment for autonomous pilot vehicles is designed to be highly open and migratable. This is so that the latest research findings and technology standards for functionally secure, low-latency, robust, and IT secure IoT communication can be quickly and flexibly integrated. At the moment, WLAN 11p represents the only car-to-x communications technology with unrestricted availability. The focus of developments is currently concentrated on application-specific functions such as WLAN 11p-based “cooperative perception” or fast video data transmission. From 2018, a standardized mobile telephony variant for car-to-x communication will be available in LTE-V. The future developments for car-to-x within 5G mobile telephony represent extensions of LTE-V. Redundant utilization of complementary communications channels in combination with a situation-adaptive choice of the optimal access technology will in future allow standards for maximum information sharing reliability and resilience to be guaranteed. Development projects on precise localization, for example for acquiring tamper-proof positioning information based on the Galileo PRS, are currently also running at Fraunhofer institutes such as IIS.

22.2.5 Roadside safety systems

For the most demanding level of automation –fully driverless vehicles – safety requirements that lie far beyond current technical achievements must be fulfilled. Autonomous vehicles need to comply with the highest security integrity level (SIL 4) and thus guarantee an error rate below 10^{-4} ... 10^{-5}. In order to at least correspond to the average safety level of a human driver, on the other hand, more than 300,000 km of accident-free driving on expressways and in highly congested urban conditions must have been demonstrated without a single intervention from the safety driver. Isolated maximums for driverless journeys without safety intervention today lie between 2,000 km and 3,000 km for Google/Waymo cars, which, however, were mostly completed on highways. Analysis of the various testing projects in California shows that automobile manufacturers are currently working at an average level of one intervention every 5 km. Based on findings from the Digitaler Knoten ("digital nodes"), Road Condition Cloud, and iFUSE projects, two strategies stand out for increasing the security of autonomous driving. Many contributions follow the traditional approach to increasing the security of autonomous driving via on-board system improvements:

- Highly integrated multisensory systems for environment recognition,
- More powerful signal and image processing,
- Learning processes for situation recognition and improving responsiveness,
- Reliable vehicle electric system infrastructure with micro-integrated electronic systems,
- Cooperative vehicle guidance, and
- Development of robust and low-latency car-to-x communication.

In addition to this, roadside safety systems based on high-performance car-to-infrastructure communications and stationary environment sensors plus object recognition, tracking, and behavior modeling are being developed. Research here focuses on

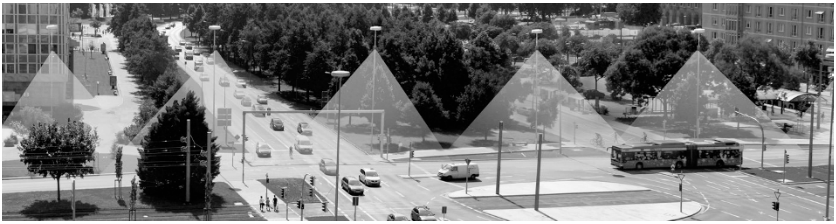


Fig. 22.1 Automation zone with roadside supervision (Fraunhofer IVI)

- Multi-level safety concepts for automation zones,
- Image-assisted safety systems in combination with stationary radar/lidar sensors,
- Cooperative environment data shared via car-to-x,
- Sensor data fusion at the perception level,
- Compression processes for external situation mapping,
- Prognosis and matching of external object motion patterns in mixed traffic environments, and
- Environment models for local weather and roadway conditions.

The roadside safety system thus transfers a defined portion of the responsibility for the safety to an intelligent infrastructure and can be understood as a kind of "virtual safety driver" in supervised automation zones.

Here, stationary environment sensors, situation prediction, and pathway planning are enhanced with information from the vehicles on recognized situations and intended trajectory. In moments of danger, the safety system coordinates the response of the supervised vehicles and resolves collision conflicts in advance or intervenes via emergency and evasive maneuvers. With the advancing development of the perceptive and predictive abilities of the vehicle guidance systems and of ever more powerful car-to-x networking, goal conflicts in external emergency interventions are increasingly avoided. Autonomous driving functions for defined scenarios outside of automation zones may be approved after a minimum number of test kilometers without external intervention. This minimum number specified is based on accident statistics from the GIDAS database. Whether roadside safety and coordination systems will remain on busy highways and intersections in future is something further development will show.

22.2.6 Digital networking and functional reliability of driverless vehicles

“Driverless vehicles” demand the highest standards in functional safety from the safety-related components implemented. Primary research goals at Fraunhofer institutes such as the ESK in Munich and IIS in Erlangen are focused on the modular, multifunctional prototyping of e/e (electric/electronic) and software architectures with fault-tolerant designs, of error-tolerant sensor and actuator equipment through to hot plug and play mechanisms. But they also primarily concentrate on the functional safety of connected e/e components and the autonomous driving functions built upon them. Over the last 10 to 12 years, the number of connected e/e systems in a vehicle has doubled to around 100 in premium cars and 80 for the mid-range. The level of connection will further increase by leaps and bounds in autonomous electric vehicles.

It is thus anticipated that modular e/e and software architectures will be key outcomes of Fraunhofer’s research. They will facilitate flexible and safe interaction of the e/e systems and of the individual interfaces and data and energy flows within the e/e systems, and variable linking of vehicle and environment data.

Sensor data can be provided by the utilization of a greater number of sensors, different physical sensor principles, and suitable sensor fusion with sufficient failure

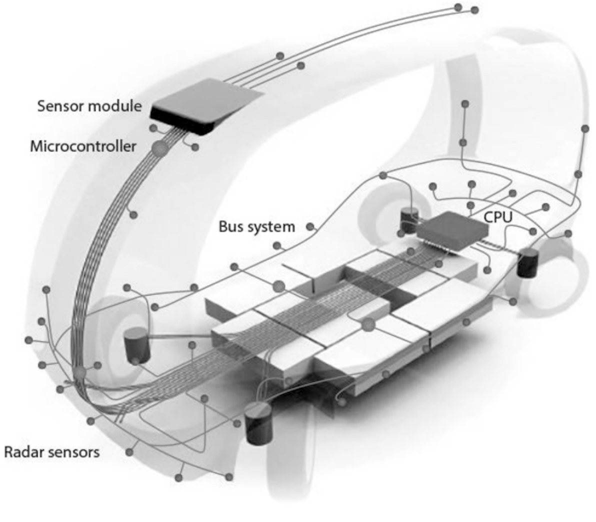


Fig. 22.2 Connected electric/electronic systems in autonomous vehicles (Fraunhofer IVI)

safety. A particularly high degree of sensor integration can be achieved through the application of extremely thin and flexible sensor systems, such as those developed at Fraunhofer EMFT, for example. Data, supply, and control paths within these systems may be selectively interconnected in the case of failure. For safety-related actuator systems, this integration scenario is significantly more complex in autonomous vehicles.

Reconfiguration and backup level management mechanisms are being developed and integrated into the software architecture and networking concept of autonomous vehicles [23] for maintaining data and energy flows in the case of the failure of individual components.

Fault tolerance and tamper proofing require demanding technical solutions even in conventional vehicle engineering. But the challenges of equipping autonomous vehicles with the ability to quickly identify and independently solve every hazardous situation inside and outside of the vehicle are significantly more complex. The particular agility of the reactions of electric vehicles and the ability of autonomous driving functions to stabilize vehicles even in extreme driving situations are addressed in a more comprehensive research scenario by Fraunhofer institutes in Dresden, Karlsruhe, Kaiserslautern and Darmstadt. Also, the extent of connected vehicles' detection horizon (far exceeding the optical field of vision in conventional vehicles) forms part of this research. From today's standpoint, a completely safe vehicle in an accident-free traffic environment still seems visionary. But autonomous driving nevertheless provides all of the technical conditions for actually reaching this goal in the not-too-distant future.

Car-to-x communication with the potential for fast ad hoc networking as well as making extensive vehicle data available (e.g. positional, behavioral, and drivetrain data; operating strategies; energy requirements; ambient data; environmental information; and traffic, car-to-x, multimedia system and interaction and monitoring data) clearly risk direct manipulation of vehicle controls and the misuse of personal or vehicle-specific data. Primary research goals at various Fraunhofer institutes within the ICT Group are thus focused on

- Fault-tolerant communication systems with corresponding coding, encryption, and data fusion,
- Security technologies with end-to-end encryption for intervehicle communications, and
- Signing and verifying sensor data.

Developing independent data rooms for industry, science, and society is also equally the subject of an extensive research and development project involving numerous Fraunhofer ICT institutes. The Industrial Data Space Association, a corporate platform founded in 2016 can now point to more than 70 member companies, including listed companies. The Fraunhofer data space reference model architecture [24] is superbly suited to ensuring data-sovereign, Internet-based interoperability as well as secure data sharing, also within intelligent traffic systems. The structure of a Germany-wide mobility data space is currently being drafted in partnership with the BMWI (Federal Ministry of Transport and Digital Infrastructure) and the German Federal Highway Research Institute (Bundesanstalt für Straßenwesen, BASt). This mobility data space is also intended to form the framework architecture for real-time supply of data to autonomous vehicles. In this way, digital mapping materials, driving time forecasts, traffic signs, etc. can be provided to autonomous driving functions in a location-referenced manner. Also, sensor information and floating car data for identifying the traffic situation can be passed back retrospectively.

22.2.7 Fast-charging capabilities and increasing ranges for autonomous electric vehicles

Combining highly precise localization with autonomous driving functions enables dynamic recharging during driving. In this process, an inductive charging strip or exact vehicle positioning is used for fully automatic conductive pulse charging via high-current underground contacts. Fast-charging processes with charging currents far exceeding 1,000 amperes at the 800-volt level cannot be handled via manual connector systems. Already in 2014, the technical foundations for superfast charging were demonstrated in practice by Fraunhofer IVI and industrial partners on Europe’s first fast-charging capable electric bus (EDDA Bus) [25]. In just five minutes, the EDDA bus can be recharged with enough power for a roughly 15 km-long bus trip using a highly automated contact system. A special traction battery was implemented with a very high power density. Currently, high-capacity batteries are briefly loadable at a charging rate of up to 5C. The availability of 10C batteries is conceivable. With these kinds of batteries, traction current for 400 km of car driving could be recharged in just 5 minutes. This nevertheless requires special autonomously passable underground charging systems. These technically demanding charging systems combined with high power-density batteries could ensure a greater spread of electromobility. It could also provide a genuinely viable alternative not only to overcoming current range limitations through higher energy densities but also by facilitating fully automated short recharging cycles with charging rates of more than 10C. Such highly stressed electric charging systems require sensor monitoring in the area close to the contact. At Fraunhofer EMFT, there are corresponding technical solutions for the abovementioned use cases – cyber physical connectors – which guarantee a sufficient level of functional safety even under conditions of extreme stress [26].

22.2.8 Vehicle design, modular vehicle construction, and scalable functionality

With the vehicle itself taking over the driving function, the construction of future automobiles will change fundamentally. The first design studies and prototypes demonstrate that the removal of engine compartments, steering wheels, and displays in sensor-driven vehicles with electric drivetrains opens up a vast creative scope. This can be used to increase flexibility, energy efficiency, and vehicle security both for passenger as well as goods transport.

The modularization strategy was already pursued in conventional automobile engineering, which meant concentrating all of the driving and control functions in a small number of base modules and completing these base modules with use case-specific body modules. The same can be implemented significantly more consistently for autonomous electric vehicles that have no driver's cockpit, which leads to a flexibilization of production, logistics, service, and utilization. This process has not even begun to be achieved in traditional automobile engineering. Even the independent exchange of body modules is possible. One example of this would be the design studies (Fig. 22.3 to 22.6) for bipart eV, an experimental electric vehicle with autonomous driving functions to VDA Level 5 ("driverless"). All of the necessary

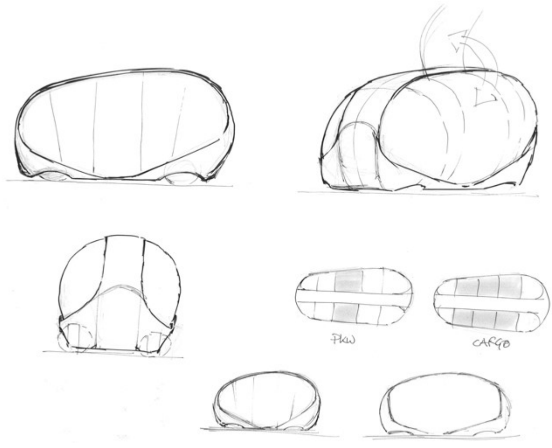


Fig. 22.3 Autonomous base module with sensor assembly in the front, roof, and rear areas, plus application-specific body modules (Fraunhofer IVI)

Fig. 22.4 bipart eV with CAP for passenger transport (Fraunhofer IVI)

components for autonomous electric driving are concentrated in the base module (Fig. 22.3). The body modules, so-called CAPs, may be offered by any manufacturer for very specific application scenarios. Economies of scale through module reuse and decoupling of the lifecycles of individual vehicle modules thus increase the efficiency and sustainability of vehicle construction enormously.

The extent to which autonomous driving functions reliably rule out individual crash load cases and thus contribute to a simplified bodywork design is a complex issue linking the mechanical structural strength of the vehicle construction with the safety of autonomous driving.

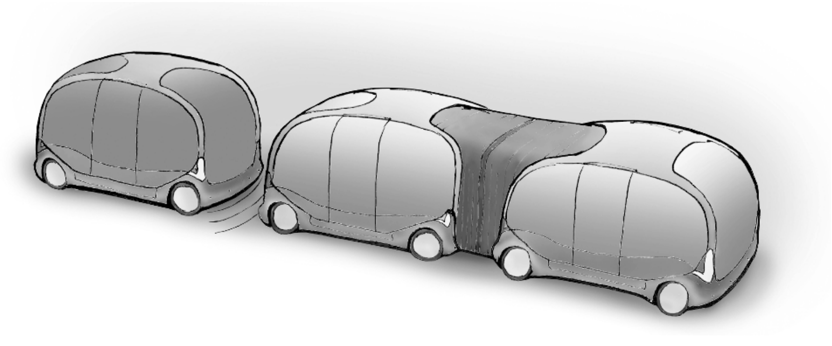


Fig. 22.5 Virtually and mechanically coupled modules (Fraunhofer IVI)

The enhanced constructive freedom in the design of autonomous vehicles incorporates the vehicle structure made up of flexibly interchangeable modules. But it is also able to be virtually or mechanically coupled (see Fig. 22.5) into multivehicle convoys that are useful primarily in logistics or in public transportation systems.

Body modules for driverless vehicles may in future be offered by manufacturers unrelated to the OEMs from a wide range of different industrial sectors. This indicates that there will be new business models and an altered competitive landscape in the automotive and utility vehicle sectors.

22.3 Autonomous transportation systems in logistics

Automated guided vehicles (AGVs) have been used within intralogistics since the 1960s. The control for early systems was based on the optical recognition of strips on the floor and later on inductive track guidance. Modern AGVs, such as those developed at Fraunhofer IML, on the other hand, move around freely in space and use hybrid odometry- and radio transmitter-based localization systems for positioning. This latest generation of transportation robots finds optimal routes to the destination of the goods independently and without track guidance.

Fig. 22.6 Multi Shuttle at Fraunhofer IML (Fraunhofer IML)

Built-in sensors enable people and other vehicles to be recognized and obstacles avoided. At present, a Smart Transport Robot (STR) for BMW production logistics has been developed, which for the first time uses automobile industry components for an AGV. The STR's energy supply, for example, is provided by recycled BMW i3 batteries. Additional standard parts from automobile production allow the inexpensive production of the suitcase-sized transport robot that, with a weight of just 135 kg, is capable of lifting, transporting, and depositing loads of up to 550 kg. The highly flexible system can thus be used in production logistics to transport dollies loaded with car parts, for example. In future, 3D camera systems will facilitate even more precise navigation and also further lower the costs of safety sensors compared with conventional AGVs.

Outside of production logistics, too, a dynamic change towards automated driving is taking place. This applies on the one hand to the automation of serial trucks – currently primarily in delimited areas (factory/yard logistics) – and on the other hand to completely new concepts for very small vehicles for use in the public sphere. Starship Technologies' delivery robot, which operates under electric power at little more than a walking pace and is primarily used for last-mile package delivery on pavements, is one such example.

Although fully automated truck journeys still require many years of testing and additional development, automation in transportation logistics nevertheless holds great potential. Alongside digitization and new drivetrain and vehicle concepts, it is one of the key drivers of change, and application scenarios are being researched at Fraunhofer IML, for example.

22.4 Driverless work machines in agricultural technology

Simple, robust tools such as plowshares, harrows, scythes, and rakes have for centuries dominated agriculture in the world's various regions. Heavy-duty tractors, complex tillage units, and fully automated sowers, fertilizer spreaders and harvesters shape our image of modern, high-productivity agricultural technology today. From a historical perspective, the development of these technologies has taken place over an extremely short timeframe and has not been without inconsistency. With increasing mechanization, greater operating widths, and rising automation, there has been a dramatic rise not only in productivity but also in machine weight. Tractors with a tractive performance of more than 350 kilowatts are able to generate turning moments of a few thousand newton meters and thus achieve tillage working widths of up to 30 m. The weight of these machines, however, stretches to 20 t and more. Moving this kind of heavy machinery over open terrain leads to high fuel requirements and to extreme compression of the soil with deep and lasting damage.

The development outlined above illustrates that sustainable agriculture that is efficient with respect to the global food situation must in future break completely new ground. This is so that highly automated agricultural machines that are as completely electrified as possible can be used in a manner that protects the soil. So-called Feldschwarm® units, which are visionary at the moment, could over the coming years prove to be a genuinely marketable migration path for future agricultural machine technology. The units are equipped with zero-emission high-efficiency drivetrain systems and work the fields as a swarm. The required pulling and work force in swarms like these is distributed electrically across the swarm vehicles' wheels and tools. Feldschwarm® units are able to move around autonomously and handle flexible working widths and variously staggered work processes as a group. Feldschwarm® technologies in lightweight design, like those currently being developed within the BMBF-sponsored Feldschwarm research project, protect the soil. Due to electrification, precise navigation, comprehensive sensor setups, and au-

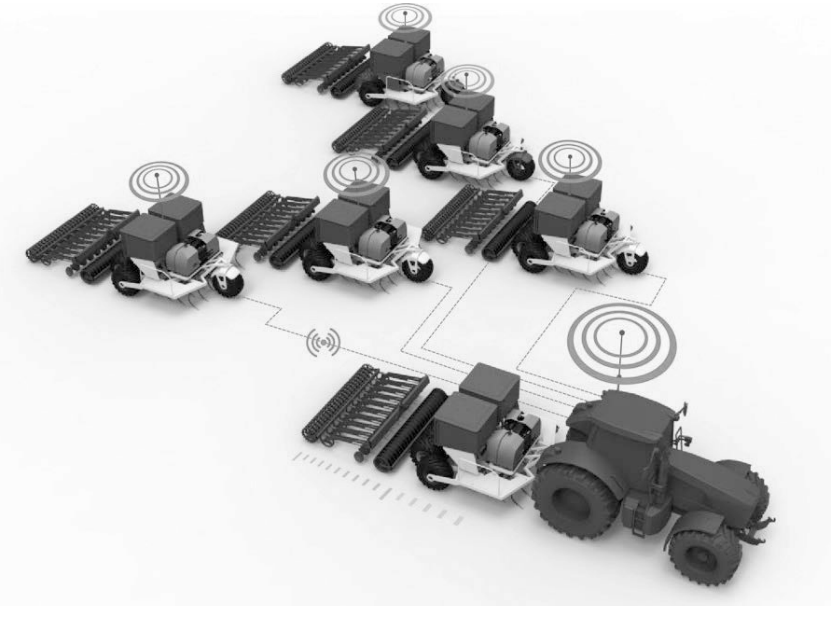


Fig. 22.7 Autonomous Feldschwarm® ("field swarm") units for tillage (© TU Dresden Chair of Agricultural Systems Technology) (Fraunhofer IVI)

tonomous driving functions, they also achieve significantly greater degrees of automation and energy efficiency than the drawn technology that is thus far still largely mechanically or hydraulically driven. Feldschwarm® units are thus destined to provide key technologies for the global development trend towards precision farming or computer aided farming. Fraunhofer, with its IVI and IWU institutes, is playing a key role in these developments.

22.5 Autonomous rail vehicle engineering

Due to the mostly external control over rail vehicle movements provided by track guidance, switch towers, and dispatchers, the automation of rail travel is far easier to solve technologically than that of road travel, and in certain cases, it is already well advanced. Driverless rail traffic between stations has been a reality in closed systems such as parts of the Nuremberg subway [27] or the Dortmund suspension railway for years. New systems (such as in Copenhagen, or lines 13 and 14 on the Paris metro [28]) are increasingly designed to be driverless. Existing subway systems are increasingly being converted to driverless operation [29]. One example, alongside the Nuremberg subway, is line 1 of the Paris metro [28].

In contrast to road automation, key control functions are still retained in central control rooms (switch towers). For regional travel, the hope is that automatic and pre-planned control will both provide improved utilization of capacity as well as contribute to energy-optimized travel. In railroad-based goods transportation in particular, however, significant costs are incurred wherever trains are joined or separated again, during switching, and where loading points are serviced. Here, significant personnel and material costs are incurred for limited ton kilometers. However, these functions are indispensable for bundling trains and servicing customers. Automation approaches are being discussed both for efficiency reasons as well as for improving workplace safety. The longevity of rolling stock and railroad infrastructure requires tailored migration concepts with, for example, a semi-automation of switching processes within factory premises and maneuvering facilities where protecting against the mistakes of other traffic participants or people is feasible. In doing so, it is entirely possible for automated guidance to be combined with conventional switching and, where relevant, even supported with sensors from the automotive sector [30], for example via radar sensors for space detection during switching. With increasing experience from semi-automated operations and additional sensor improvements, the utilization of driverless operated vehicles over regional journeys is being considered as a next step.

22.6 Unmanned ships and underwater vehicles

More than 70% of the earth's surface is water. Ocean-going ships often travel for many weeks, requiring crews with specific nautical competencies. As recently as 2014, the production of autonomously guided ships was viewed as highly unlikely by 96% of German ship owners. Just two years later, 25% were convinced that unmanned shipping was possible (PwC, ship owners study, issue 2014 and 2016). Demanding operating environments on the high seas, the restrictiveness of data transmission, and constant monitoring of the technical systems on board, alongside unanswered legal questions made autonomous shipping unlikely so far. But industrial companies, classification societies, and research institutes (not least Fraunhofer CML in Hamburg) have in the meantime made promising progress. Due to improvements in and the acceleration of data transmission, continuing digitization, and the development of specific solutions for autonomous shipping, the future realization of this vision appears conceivable. More recent studies have focused on ships that are intended for autonomous use on fixed routes rather than for global use. This is because present conditions make constant digital supervision and control more possible. The areas of application selected are suitable for low-maintenance electric or hybrid propulsion. In Norway, for example, an electric container feeder ship for distances of up to 30 nautical miles is currently being developed that is due to operate autonomously from 2020 following a transition phase.

Fig. 22.8 Visualization of an autonomous evasive maneuver and weather routing (Fraunhofer CML)

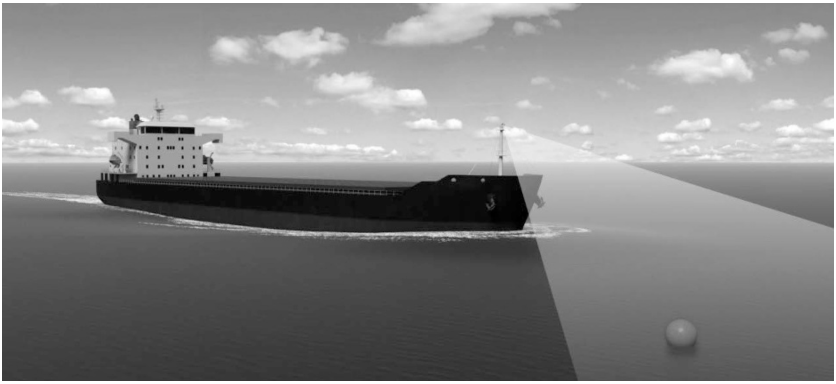


Fig. 22.9 Visualization of an automated crow's nest – object recognition via camera systems (Fraunhofer CML)

Fraunhofer CML has been working intensively with solutions for unmanned and autonomously guided commercial ships in recent years: first under the EU-sponsored MUNIN research project [31], and then in partnership with South Korea's Daewoo Shipbuilding and Marine Engineering DSME. Three ship guidance simulators are being utilized here, e.g., in the ongoing development of systems for independently conducting evasive maneuvers. Autonomously guided ships must also operate in accordance with international collision avoidance regulations in order to avoid dangerous situations and collisions. Their steering must thus respond at least as reliably as would be expected from a human helmsman, in keeping with the regulations. The solution developed at Fraunhofer CML is able to combine data from different sensor sources such as radar, the automatic identification system (AIS), or day- and night-vision cameras and produce a picture of the traffic situation. Should the situation require it, an evasive maneuver based on the international collision prevention rules is computed and executed. The simulation environment is supplemented by a Shore Control Center (SCC). The SCC permits the monitoring and control of a fleet of autonomously navigating ships from the shore. Generally, these fleets operate on the high seas without relying on external support. Should, however, automated on-board systems become overwhelmed by a situation, the Shore Control Center is able to intervene instantly. In this way, by increasing efficiency and safety on board, the development of assistance systems for commercial shipping is already making important contributions to the transition towards autonomous shipping.

Vehicles also navigate underneath the water's surface in order to produce underwater maps, investigate marine sediments, and carry out inspections and measurements. Areas of application, alongside marine research, include preparations for laying deep sea cables and pipelines and their inspection [32].

Nevertheless, neither radio waves nor optical or acoustic signals permit continuous communication under water. DEDAVE, a flexible deep-sea underwater vehicle by Fraunhofer IOSB, is able to dive to depths of 6,000 m, to conduct missions lasting up to 20 hours, and is equipped with numerous sensors. Due to the patented fast-switching mechanism for batteries and data stores, maintenance and setup times are shorter, reducing the duration and costs of a mission significantly. Current research studies are investigating the operation of a swarm of 12 intelligent deep-sea robots [33].

Sources and literature

[1] IHS Automotive (2014): Emerging technologies. Autonomous cars – Not if, but when

[2] Victoria Transport Institute (2013): Autonomous vehicle implementation predictions. Implications for transport planning. http://orfe.princeton.edu/~alaink/SmartDrivingCars/Reports&Speaches_External/Litman_AutonomousVehicleImplementationPredictions.pdf [letzter Zugriff: 03.07.2017]

[3] Strategy& (2014): Connected Car Studie 2014

[4] O. Wyman Consulting (2012): Car-IT Trends, Chancen und Herausforderungen für IT- Zulieferer

[5] Zentrum für europäische Wirtschaftsforschung ZEW, Niedersächsisches Institut für Wirtschaftsforschung (NIW) (2009): Die Bedeutung der Automobilindustrie für die deutsche Volkswirtschaft im europäischen Kontext. Endbericht an das BMWi

[6] McKinsey (2014): Connected car, automotive value chain unbound. Consumer survey

[7] Appinions (2014): Autonomous cars. An industry influence study

[8] J.R. Ziehn (2012): Energy-based collision avoidance for autonomous vehicles. Masterarbeit, Leibniz Universität Hannover

[9] M. Ruf, J.R. Ziehn, B. Rosenhahn, J. Beyerer, D. Willersinn, H. Gotzig (2014): Situation Prediction and Reaction Control (SPARC). In: B. Färber (Hrsg.): 9. Workshop Fahrerassistenzsysteme (FAS), Darmstadt, Uni-DAS e.V., S. 55-66

[10] J. Ziegler, P. Bender, T. Dang, and C. Stiller (2014): Trajectory planning for Bertha. A local, continuous method. In: Proceedings of the 2014 IEEE Intelligent Vehicles Symposium (IV), S. 450-457

[11] J.R. Ziehn, M. Ruf, B. Rosenhahn, D. Willersinn, J. Beyerer, H. Gotzig (2015): Correspondence between variational methods and hidden Markov models. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV), S. 380-385

[12] M. Ruf, J.R. Ziehn, D. Willersinn, B. Rosenhahn, J. Beyerer, H. Gotzig (2015): Global trajectory optimization on multilane roads. In: Proceedings of the IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), S.1908-1914

[13] J.R. Ziehn, M. Ruf, D. Willersinn, B. Rosenhahn, J. Beyerer, H. Gotzig (2016): A tractable interaction model for trajectory planning in automated driving. In: Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), S. 1410-1417

[14] N. Evestedt, E. Ward, J. Folkesson, D. Axehill (2016): Interaction aware trajectory planning for merge scenarios in congested traffic situations. In: Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), S. 465-472

[15] T. Emter und J. Petereit (2014): Integrated multi-sensor fusion for mapping and localization in outdoor environments for mobile robots. In: Proc. SPIE 9121: Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications

[16] T. Emter, A. Saltoğlu, J. Petereit (2010): Multi-sensor fusion for localization of a mobile robot in outdoor environments. In: ISR/ROBOTIK 2010, Berlin, VDE Verlag, S. 662- 667

[17] C. Frese, A. Fetzner, C. Frey (2014): Multi-sensor obstacle tracking for safe human-robot interaction. In: ISR/ROBOTIK 2010, Berlin, VDE Verlag, S.784-791

[18] C. Herrmann, T. Müller, D. Willersinn, J. Beyerer (2016): Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs. In: Proc. SPIE 9987, Electro-Optical and Infrared Systems: Technology and Applications

[19] J. Uhrig, M. Cordts, U. Franke, T. Brox (2016): Pixel-level encoding and depth layering for instance-level semantic segmentation. 38th German Conference on Pattern Recognition (GCPR), Hannover, 12.-15. September

[20] C. Frese und J. Beyerer (2011): Kollisionsvermeidung durch kooperative Fahrmanöver. In: Automobiltechnische Zeitschrift ATZ Elektronik, Jg. 6, Nr. 5, S. 70-75

[21] C. Frese (2012): Planung kooperativer Fahrmanöver für kognitive Automobile. Dissertation, Karlsruher Schriften zur Anthropomatik, Bd. 10, KIT Scientific Publishing

[22] M. Düring, K. Lemmer (2016): Cooperative maneuver planning for cooperative driving. In: IEEE Intelligent Transportation Systems Magazine, Bd. 8, Nr. 3, S. 8-22

[23] J. Boudaden, F. Wenninger, A. Klumpp, I. Eisele, C. Kutter (2017): Smart HVAC sensors for smart energy. International Conference and Exhibition on Integration Issues of Miniaturized Systems (SSI), Cork, 8.-9. März

[24] https://www.fraunhofer.de/en/research/lighthouse-projects-fraunhofer-initiatives/industrial-data-space.html [letzter Zugriff: 03.07.2017]

[25] http://www.edda-bus.de/ [letzter Zugriff: 03.07.2017]

[26] F.-P. Schiefelbein, F. Ansorge (2016): Innovative Systemintegration für elektrische Steckverbinder und Anschlusstechnologien. In: Tagungsband GMM-Fb. 84: Elektroni- sche Baugruppen und Leiterplatten (EBL), Berlin, VDE Verlag; F. Ansorge (2016): Stei- gerung der Systemzuverlässigkeit durch intelligente Schnittstellen und Steckverbinder im Bordnetz. Fachtagung Effizienzsteigerung in der Bordnetzfertigung durch Automa- tisierung, schlanke Organisation und Industrie-4.0-Ansätze, Nürnberg, 5. Oktober

[27] G. Brux (2005): Projekt RUBIN. Automatisierung der Nürnberger U-Bahn. Der Eisenbahningenieur, Nr. 11, S. 52-56

[28] http://www.ingenieur.de/Branchen/Verkehr-Logistik-Transport/Fahrerlos-Paris-Die-Metro-14-um-sechs-Kilometer-verlaengert [letzter Zugriff: 03.07.2017]

[29] A. Schwarte, M. Arpaci (2013): Refurbishment of metro and commuter railways with CBTC to realize driverless systems. In: Signal und Draht, Jg. 105, Nr. 7/8, S. 42-47

[30] Verband der Automobilindustrie e.V. (2015): Automatisierung Von Fahrerassistenzsystemen zum automatisierten Fahren; Institute for Mobility Research (2016): Autonomous Driving The Impact of Vehicle Automation on Mobility Behaviour

[31] http://www.unmanned-ship.org/munin/ [letzter Zugriff: 03.07.2017]

[32] https://www.fraunhofer.de/content/dam/zv/de/presse-medien/Pressemappen/hmi2016/Presseinformation%20Autonome%20Systeme.pdf [letzter Zugriff: 03.07.2017]

[33] https://arggonauts.de/de/technologie/ [letzter Zugriff: 03.07.2017]