



AI5production - SYMPOSIUM

Digitalization in production – solutions and applications

Results from three years
EDIH AI5production

15. OCTOBER 2025

JKU LINZ, OIC STUFENFORUM



Co-funded by
the European Union



European
Digital Innovation
Hubs Network



Bundesministerium
Wirtschaft, Energie
und Tourismus



FFG
Forschung wirkt.



Editor

Johannes Kepler Universität Linz
Altenberger Straße 69
4040 Linz

editorial management

Univ.-Prof.in Dr.in Barbara Krumay, Bakk MSc(WU)
barbara.krumay@jku.at

The authors of the articles are responsible for the accuracy of their content!

Linz, 15.10.2025

Foreword

Foreword Dipl.-Ing. Franz Sümeecz

TU Wien Pilotfabrik, Consortium Leader EDIH AI5production



The weak economy, US trade policy, and the decarbonization of industry pose enormous challenges for manufacturing companies in Europe, and in Austria in particular. For companies, it is a matter of being competitive and, above all, resilient in a global market. Digitalization and, above all, AI play a central role in this. Austrian universities and non-university research institutions are at the forefront of international research and innovation. The aim of the EDIH is to promote technology transfer in order to bring existing know-how and expertise to bear in manufacturing companies, thereby providing sustainable support for the Austrian economy. Digital technologies enable increased efficiency in production processes and workflows. The aim is to reduce costs, increase sales, and become more competitive and resilient. EDIH AI5production shows how

digitalization and AI contribute to this. With more than 150 projects with companies in the last three years, 50% of which involved AI methods, EDIH is an Austrian success story. The simple, low-threshold, and fast support that the EDIH offers companies and the extensive expertise of the consortium consisting of 17 institutions are essential. Projects that demonstrate the potential of AI were selected for this symposium.

Univ.-Prof.in Dr.in Barbara Krumay, Bakk. MSc(WU)

JKU Business School, Lead of Work Package „Support to find Investment“



The EDIH AI5production has succeeded in driving forward digitalization in SMEs and mid-caps in Austria. The diverse contributions submitted by the partners of EDIH AI5production impressively demonstrate this. The projects that form the basis for the contributions are only the beginning. The focus is not on individual solutions, but on the generalizability of the findings as a contribution to science and the economy. This results in practical examples (“use cases”) that have been developed on a scientific basis and serve as inspiration for further digitalization projects. Especially for the use and implementation of AI applications, the call for well-documented use cases is growing louder in order to protect companies from bad investments and to be able to fully exploit the potential of AI. The

examples summarized in this volume are part of the AI5production symposium, which will take place on October 15, 2025, at the JKU in the OIC Stufenforum. Three of the contributions deal with the role of EDIH AI5production in increasing the digital maturity of SMEs and mid-caps in the EU. The remaining contributions illustrate, at various technical levels, how digitalization can be implemented in manufacturing companies. Even though AI appears in many contributions and has been used in many TBIs, other established digital technologies must not be forgotten in order to advance the digital maturity of European and Austrian SMEs and mid-caps.

Table of contents

Digital maturity of small and medium-sized enterprises through innovation impulses.....	3
Training and corporate culture as factors for AI integration.....	9
Benefits for manufacturing companies through the European innovation ecosystem.....	14
Artificial intelligence in production	20
Digital data consistency for drive technology in building services engineering	28
Efficiency improvements in WIG and plasma welding inverters	32
Multi-sensor imaging for the qualitative characterization of used solar modules	37
Digitization of an energy efficiency concept with reporting function	40
Casting process for sand molds with dynamic vacuum control ROBOCAST V.....	45
CFD simulation as a digital service for development and production	49
Intelligent chatbots for querying company knowledge- ALLEX	56
Optimization of forming processes through AI-based monitoring and control	63
Digital human-centered assistance systems	71
Intelligent assistance system for flexible human-robot collaboration in manufacturing	77
Open source cybersecurity anomaly detection.....	82
RAG is Not All You Need!	87
Reconstruction pipeline for low-cost images for creating digital twins	93
Towards Optical Tracking for AR Assisted Oral Surgery	101

Glossary

- European Digital Innovation Hub (EDIH) AI5production – one of four hubs in Austria funded by the EU and FFG with a focus on manufacturing SMEs and mid-caps
- National Digital Innovation Hubs (DIHs) – regional contact points for digitalization issues.
- Test before Invest (TBI) – small-scale project on the topic of digitalization, carried out by a partner of EDIH AI5production for a company. A TBI includes two training sessions (technology training at the beginning of the application and “on-site training” at the end of the project).
- Support to find investment (STFI) – services and consulting on financing digitization projects within the framework of EDIH AI5production. The spectrum ranges from consulting with regard to further research funding (in case further developments are necessary) to support in the search for venture capital or business angels.
- SMEs & mid-caps – according to the European Union definition with a focus on the number of employees.
- Start-ups – companies with innovative business ideas and models up to a maximum of 5 years after their founding
- Artificial intelligence (AI) – generally understood as non-rule-based technology that mimics human intelligence – further definitions and categories can be found in the articles

Digital maturity of small and medium-sized enterprises through innovation impulses

Barbara Krumay
JKU Business School, Johannes Kepler
Universität Linz
barbara.krumay@jku.at

Manuel Mühlburger
JKU Business School, Johannes Kepler
Universität Linz
manuel.muehlburger@jku.at

Kurzzusammenfassung

Für kleine und mittlere Unternehmen (KMUs) ist die Umsetzung von Digitalisierungsprojekten und damit einhergehende die Hebung der Digitalen Reife herausfordernd. Es fehlen oft die finanziellen und personellen Möglichkeiten. Darüber hinaus müssen auch interne Voraussetzungen erfüllt werden, damit Digitalisierung möglich ist. In der Regel sind kleiner Unternehmen auch vorsichtig bei Investitionen in neue Technologie, daher fehlt oft eine Infrastruktur zum Testen neuer digitaler Technologien. Externe Innovationsimpulse sind für KMUs dann von großer Bedeutung, wenn sie leicht zugänglich und unabhängig (z.B. von Herstellern) sind. Basierend auf Interviews mit Vertreter:innen von Unternehmen, die die Services des EDIH AI5production in Anspruch genommen haben, wurde festgestellt, dass diese Innovationsimpulse zur Umsetzung von Digitalisierungsprojekten geführt haben und dadurch digitalen Reifegrad gesteigert werden konnte.

Keywords: Digitalisierung, Digitale Reife, Innovationsbereitschaft, Innovationsimpulse

Abstract

The implementation of digital technologies and digitalization projects is challenging, in particular for small and medium-sized enterprises (SMEs). They often lack the financial and human resources to implement such projects. In addition, digitalization projects also require the fulfillment of internal preconditions. Furthermore, SMEs are generally cautious about investing in new technology and often lack the infrastructure to test new digital technologies. External innovation impulses are of great importance for SMEs if they are easily accessible and independent (e.g., from manufacturers). Based on interviews with companies that have been consumed the services of the EDIH AI5production, it has been shown that these innovation impulses can contribute to implementing digitalization projects and raising the level of digital maturity.

Keywords: Digitalisation, digital maturity, Innovation readiness, innovation impulses

Introduction

Small and medium-sized enterprises account for 99% of European businesses and are often referred to as the driving force behind the European economy [1]. The European Union (EU) has therefore launched initiatives to support SMEs in their digitalisation projects and raise their digital maturity [1]. Many challenges can also be traced back to the development of SMEs, such as historically grown structures and unclearly defined responsibilities – many issues are elevated to ‘top management issues’, regardless of whether the relevant expertise is available. Many SMEs are run by their founders or owners – often they are also family businesses [2], which are characterised by the fact that strategic decisions are primarily made by the founders and family members. These decision-makers have experience in the company's core business, but often lack the time to build up technical expertise. This also has an impact on the companies' willingness to embrace digitalisation. Family-run companies also often show a low willingness to take risks [2], which primarily hinders innovation. Governance structures that regulate responsibilities exist only implicitly in the true sense of the word, i.e. knowledge about responsibility for decisions is not explicitly defined and documented within the company [2].

SMEs are defined in the EU by the number of employees, turnover and balance sheet total. Medium-sized enterprises therefore have a maximum of 250 employees, a turnover of max. €50 million and a balance sheet total of max. €43 million. For small and micro-enterprises, total assets and turnover are the same, i.e. small enterprises have a maximum of 50 employees and a turnover of €2 million, while micro-enterprises have a maximum of 10 employees and a turnover of €2 million [3]. In addition, the EU defines so-called mid-caps with up to 3,000 employees [4].

The size of SMEs gives rise to certain characteristics (e.g. scarcity of resources, lack of research and development, goals driven by operational business, no documented corporate strategy) that can limit their innovative strength. Companies in which founders predominantly hold decision-making positions [5] are usually younger companies in the start-up phase. SMEs that were founded no more than five years ago and are experiencing growth in terms of employees or turnover thanks to innovative technology or an innovative business model are referred to as start-ups. In reality, start-ups are usually founded by an individual or a team of founders who share a common goal or idea. In addition to start-ups, the European SME landscape includes many family businesses that share certain characteristics, regardless of the size and history of the company. These include the fact that decisions are made by the members of the family who founded the company, at least one family member is involved in the management of the company, and more than one family member holds a decision-making position in the company [7], [8], [9]. One-person companies are therefore usually excluded from consideration [7], [8]. Family businesses are characterised by unique human resources, specific structures and succession arrangements [8]. The success of family businesses can be influenced both positively – through stronger identification with the company – and negatively – through the family situation and financial dependence – by the involvement of family members [10]. All companies go through a maturation process and, as a result, a structural transformation. This can be triggered by both external (e.g. market situation, new technologies) and internal influences (e.g. change at the top of the company). Every transformation brings about changes that depend on various factors (e.g. corporate culture, willingness to change and innovate). The digital transformation of companies is a transformation driven by existing universal digital technologies and their adoption in the market. In order to enable the digital transformation of SMEs, digital skills and innovative capacity (as internal factors) as well as an adequate environment (external factors) must be in place [6]. Due to limited human and financial resources, external impetus is also required for SMEs to achieve this transformation. External innovation impetus can take various forms, but these offerings must be designed to be as accessible as possible. In addition to the opportunity to test the use of new technologies (TBI), these stimuli also include financial instruments (e.g. in the context of STFI funding, contact with business angels), training opportunities and networking opportunities. This article examines how external innovation stimuli can contribute to the implementation of digitisation projects and thus to increasing digital maturity.

Procedure

In order to investigate how external innovation impulses can increase the digital maturity of SMEs, interviews were conducted with three companies that carried out a TBI as part of the EDIH AI5production. The companies were deliberately selected so that they differed in terms of their internal factors, while the innovation stimulus provided by the TBI was always approximately the same (use of EDIH AI5production services, documented increase in digital maturity). A family business (company A), a start-up (company B) and a company that had already undergone restructuring (company C), i.e. had undergone a transformation process, were selected (see Table 1). Between May and July 2025, guided interviews were conducted with several people in the company. These interviews were recorded, transcribed using AI and analysed qualitatively.

Table 1. Summary of the companies & interviews

	Company A	Company B	Company C
Classification	Family business	Start-Up	Restructured
Age in years	~ 25	< 5	> 50
Number of employees	< 250	< 50	< 3.000
Business area	metalworking	supplier	electronics
Technology	AI application	Platform	AI application
Interview partner	Managing director (A1) Production manager (A2)	Founder (B1 & B2)	Managing director (C1) Innovation manager (C2) Production employee (C3)

The interview guide was developed from the literature and contained four key questions:

- What external innovation impulses do SMEs need in order to implement digitalisation in their companies? (Demand)
- How must external innovation impulses be designed so that they are accepted by SMEs? (Design)
- What internal factors influence the success of innovation impulses? (Internal factors)
- What examples are there within the company of how external innovation impulses have contributed to increasing digital maturity? (Examples)

In addition, short introductory questions (about the company and the person) were asked. During the interview, questions from the interviewees regarding comprehension were addressed. A conscious decision was made not to mention EDIH AI5production in the key questions. The interviews lasted an average of 47 minutes.

Results

In order to illustrate how external innovation impulses can contribute to increasing the digital maturity of SMEs, the results of the interviews are presented on the basis of the key questions. Some statements by interview partners are paraphrased and marked with the initials of the interview partner.

With regard to the first key question, ‘What external innovation impulses do SMEs need to implement digitalisation in their companies?’, a wide variety of possibilities were mentioned, which can be summarised in three categories: monetary aspects, know-how development and access to technologies. Among the monetary innovation impulses, subsidies and access to investors were mentioned in particular. The interviewees from Company B (start-up) pointed to the large but often confusing subsidy landscape in Austria and also mentioned business angels (B2) as an important innovation impulse, since ‘business angels only take innovative companies under their wing’ (B2). Interviewees A1 and A2 pointed out the strong position of the company's main bank, which can also stimulate innovation by providing appropriate loans. However, it was also said that ‘five years ago, it was even easier to obtain loans for innovative projects’ (A1). The category ‘know-how’ subsumed all innovation stimuli related to training and education. While ‘self-learning via YouTube’ predominates in company B, the interviewees from the other two companies have a very clear idea of the need for structured offerings. Interviewee A2 said that ‘continuous training on new technologies, but also refreshing know-how on technologies already in use, is absolutely necessary.’ Interviewee C3 pointed out that ‘AI cannot be mastered through on-the-job training alone’. Possible forms included hackathons, boot camps and training via chatbots. All interviewees emphasised that training must go hand in hand with the introduction of new technologies or, ideally, begin well in advance of their introduction (C1). Only then can training generate momentum for innovation. Access to technology that is not available in the company was also mentioned as another external innovation impulse. In particular, it was pointed out that the high investment costs are not reasonable for many technologies. The interviewees from companies A and B particularly pointed out that only companies above a certain size can reasonably

pursue their own research and development efforts beyond the normal development of their core business. Interviewee B1 said: 'We use all possible access to hackerspaces, factories and the like to try out new technologies – we cannot afford this infrastructure ourselves' (B1).

All interviews pointed out that external innovation impulses played an important role in the implementation of digitisation projects. It was mentioned several times that some projects within the company were only possible thanks to external innovation impulses. The role of EDIH AI5production was also directly addressed, with one respondent noting that 'we certainly would not have started this project without AI5production' (C3). TBIs were explicitly referred to as the initial spark for a later project (A1, C2). STFI was also specifically mentioned as an important prerequisite and a crucial service for implementation (B1, B2, C1). The service was rated as particularly important and valuable. Interestingly, the Digital Maturity Assessments, which must be carried out before and after each TBI, were also mentioned. These 'give a good overview of where we stand' (A2).

When asked how external innovation initiatives should be designed in order to be accepted by SMEs, there was a clear answer that can best be summarised by the term 'low threshold'. It is striking that the answers were mainly negative in connotation (e.g. 'not complicated', 'no long applications', 'little bureaucracy', 'not hidden and only for experts'). Interviewee B2 said that 'as a start-up, you have many advantages and benefit from many services – other SMEs often don't know about them'. Interviewee C2 said, 'once you're in this cycle [sic], it gets easier – getting in is difficult' (C2). The interviewees from companies said that they had little or no experience with this so far, but had quite clear ideas about how external innovation impulses could be 'made available to us too' (A1) – these include tailor-made offers ('so you don't have to rummage through them all') and short-term opportunities. ('must fit into our business schedule'). Costs were addressed in the sense that external innovation impulses must be 'free or reasonable' (B2) in order to be accepted. It was also emphasised that all measures 'require a lot of effort', which is also reflected in the commitment of resources and thus also incurs costs. Interviewee B2 commented that 'you put a lot of time and effort into an application, which is then rejected.' In line with this, all interviewees mentioned a kind of assessment of the chances of success when external innovation impulses (especially funding, access to technology) are considered by companies. Another important factor for the acceptance of external innovation impulses is their independence. All interviewees stated that pure 'sales events' by technology providers are not understood as impulses but rather as 'aggressive marketing' (B2) and are therefore often rejected.

The question of internal factors that determine the success of external innovation impulses was also answered very clearly. All interviewees agreed that an open corporate culture and support from management are two basic prerequisites. Other factors mentioned were employees' willingness to embrace digitalisation and their current level of knowledge (A1, B1, B2, C1), a flexible or ambidextrous (combination of stability and agility) structure (B1, C1, C2, C3) and the reduction of fears about change (A1, B1, C2). The development of role models ('champions' / 'role models' / 'stewards') was also addressed in various ways. These role models can be used to initiate a targeted and honest exchange of experiences, which provides clarity about external innovation impulses and how to use them. This can also help to eliminate fears and uncertainties in times of change. The final question in the guide asked about examples within the company of how external innovation impulses had contributed to increasing digital maturity. Although the EDIH AI5production services were deliberately not mentioned, all interviewees cited the services (TBIs, STFI, training courses) as examples. The low threshold of the services was highlighted, as was 'the understanding of all those involved in EDH AI5production of the challenges in the manufacturing sector' (A1). Individual and direct communication and the comprehensive service culture were also mentioned. Other examples such as general cooperation with research institutions (universities, technical colleges, research institutes), the services of various providers (Chamber of Commerce, Austrian Business Service) and the opportunities to 'learn from the experience of others' (B2) were also mentioned.

Interestingly, networking was not mentioned as an influencing factor in any of the interviews. When asked, time and 'fit' were mentioned above all else as factors that prevent more intensive networking in

the context of external innovation impulses. ‘Fit’ refers both to content (networking on a specific topic) and to the company. Interviewee A2 commented, ‘We already know who suits us’ (A2). However, two interviewees (B1, C2) also mentioned the positive effect of well-designed and appropriately orchestrated networking events and identified them as an external innovation stimulus.

In addition to the guide questions, an analysis was also conducted to determine whether the classification of companies as family businesses, start-ups and restructured companies led to different results. In addition to a quantitative-interpretative analysis, emotions expressed during the interviews were also used for the assessment. The mentions were divided into 0 / and + - +++ to represent the frequency of the mention, assessed in relation to each other and averaged across the interview partners per company. Emotionality on a specific topic is operationalised with * (no star – no emotion, * - *** from low to high emotionality). The first three questions of the guide were examined.

Tabel 2 Summary of the companies & interviews

	company A	company B	company C
Classification	Family business	Start-Up	Restructured
Requirements			
Monetary aspects	+++ / ***	+++ / *	+ / 0
Know-how development	++ / *	++ / **	+++ / *
Access to technologies	+++ / **	+++ / ***	+ / 0
Design			
Low threshold	++ / ***	+++ / 0	++ / *
Internal factors			
Open corporate culture	+ / 0	++ / **	+++ / *
Support from management	++ / **	+++ / 0	+++ / ***

Conclusion

External innovation impulses are predominantly seen as positive and necessary for implementing digitisation projects and thereby increasing the digital maturity of SMEs and mid-caps. In addition to monetary aspects, know-how development and access to technologies were cited as important innovation impulses. These must be designed to be as low-threshold as possible so that they can also be utilised. The companies themselves must create the conditions necessary for the innovation impulses to be reflected in increased digital maturity. Positive examples of external innovation impulses show that, in addition to low thresholds, an understanding of the situation of manufacturing companies also plays an important role. It also became clear that independent offerings are preferred over manufacturer-specific ones. It was also found that the interviewees valued external innovation drivers, in particular the services of the EDIH. AI5production, which are crucial for the implementation of digitisation projects. A comparison of the results in terms of the classification of the companies shows that companies A and B have many similarities. The focus is on the monetary aspects and the possibility of accessing technologies. Contrary to the frequently expressed peculiarity of family businesses, this does not seem to be true for innovation impulses. It is striking that emotionality was very high in the family businesses surveyed when it came to monetary aspects. Both interview partners reported very emotionally on their unsuccessful attempts to obtain funding. Company C – as a mid-cap and the largest and oldest company in the study – appears to be very balanced in all aspects. Emotionality could only be identified in Company C with regard to management support. Looking at the statements on examples of successful innovation impulses, one could also conclude that offers for mid-caps should be designed differently in order to increase digital maturity. However, no general conclusions can be drawn from the three companies examined.

References

- [1] Europäisches Parlament, „Kleine und mittlere Unternehmen | Kurzdarstellungen zur Europäischen Union | Europäisches Parlament“, Kurzdarstellung zur Europäischen Union. Zugegriffen: 5. August 2025. [Online]. Verfügbar unter: <https://www.europarl.europa.eu/factsheets/de/sheet/63/kleine-und-mittlere-unternehmen>
- [2] M. R. W. Hiebl, „Risk aversion in the family business: the dark side of caution“, *Journal of Business Strategy*, Bd. 35, Nr. 5, S. 38–42, Sep. 2014, doi: 10.1108/JBS-09-2013-0087.
- [3] „Kleine und mittlere Unternehmen - EUR-Lex“. Zugegriffen: 14. August 2025. [Online]. Verfügbar unter: <https://eur-lex.europa.eu/DE/legal-content/glossary/small-and-medium-sized-enterprises.html>
- [4] „SME definition - European Commission“. Zugegriffen: 14. August 2025. [Online]. Verfügbar unter: https://single-market-economy.ec.europa.eu/smes/sme-fundamentals/sme-definition_en
- [5] B. Hendricks, T. Howell, und C. Bingham, „How much do top management teams matter in founder-led firms?“, *Strategic Management Journal*, Bd. 40, Nr. 6, S. 959–986, Juni 2019, doi: 10.1002/smj.3006.
- [6] C. F. A. Arranz, M. F. Arroyabe, N. Arranz, und J. C. F. de Arroyabe, „Digitalisation dynamics in SMEs: An approach from systems dynamics and artificial intelligence“, *Technological Forecasting and Social Change*, Bd. 196, S. 122880, Nov. 2023, doi: 10.1016/j.techfore.2023.122880.
- [7] C. Dörflinger, A. Dörflinger, K. Gavac, und B. Vogl, „Endbericht Familienunternehmen“, *KMU Forschung Austria*, Wien, 2013.
- [8] J. Diéguez-Soto, P. López-Delgado, und A. Rojo-Ramírez, „Identifying and classifying family businesses“, *Rev Manag Sci*, Bd. 9, Nr. 3, S. 603–634, Juli 2015, doi: 10.1007/s11846-014-0128-6.
- [9] Europäische Kommission, „Overview of family business relevant issues“, Europäische Kommission, 30-CE-0164021/00–51, 2009. [Online]. Verfügbar unter: http://ec.europa.eu/enterprise/policies/sme/promoting-entrepreneurship/family-business/index_en.htm
- [10] A. Stewart und M. A. Hitt, „Why Can’t a Family Business Be More Like a Nonfamily Business? Modes of Professionalization in Family Firms“, *Family Business Review*, Bd. 25, Nr. 1, 2012.
- [11] B. Weber-Lewerenz, „Corporate digital responsibility (CDR) in construction engineering—ethical guidelines for the application of digital transformation and artificial intelligence (AI) in user practice“, *SN Applied Sciences*, Bd. 3, Nr. 10, 2021, doi: 10.1007/s42452-021-04776-1.

Training and corporate culture as factors for AI integration

Karin Koivisto
Pilotfabrik TU Wien – Industrie 4.0
karin.koivisto@tuwien.ac.at

Kurzzusammenfassung

Künstliche Intelligenz (KI) in der Produktion ist nicht nur eine technologische Herausforderung, sondern brauchen in vielen Fällen organisationale Veränderung um erfolgreich zu sein. Um KI in produzierenden Unternehmen erfolgreich einsetzen zu können, müssen die Mitarbeiter:innen entsprechend geschult werden. Neben der Vermittlung von technologischen Inhalten können diese Schulungen aber auch, eine organisationale Veränderung vorantreiben. Neben den technologischen Themen wie grundlegendes Verständnis von KI, praxisnahe Anwendungsbeispiele sowie der Aufbau von Datenkompetenz sollen Mitarbeitende auch befähigt werden, Prozesse ganzheitlich zu betrachten und Potenziale entlang der Wertschöpfungskette zu identifizieren. Die Ergebnisse aus dem EDIH AI5production haben gezeigt, dass praxisorientierte Anwendungsfälle und integrative Workshops hinaus Akzeptanz, Digitale Reife und Bereitschaft zum Wandel fördern. Die Unternehmenskultur muss sich dafür aber in Richtung offene Lern- und Fehlerkultur entwickeln und die aktive Partizipation der Mitarbeiter:innen fördern, um Ängste abzubauen. Durch KI-Schulungen können somit nicht nur technische Fähigkeiten geschult werden, sondern darüber hinaus auch die Unternehmenskultur verändert werden, um einen nachhaltigen KI-Einsatz möglich zu machen.

Keywords: Künstliche Intelligenz (KI), Produktion, Mitarbeiterschulung, Unternehmenskultur

Abstract

Artificial intelligence (AI) in production is not only a technological challenge, but in many cases also requires organisational change in order to be successful. AI implementation in manufacturing companies requires training for the employees. In addition to technological content, these training courses can also drive organisational change. Besides pure technological topics such as a basic understanding of AI, practical application examples and the development of data literacy, employees should also be empowered to take a holistic view of processes and identify potential along the value chain. The results from EDIH AI5production have shown that practice-oriented use cases and integrative workshops promote acceptance, digital maturity and willingness to change. However, the corporate culture must evolve towards an open learning and error culture and encourage active participation by employees in order to reduce fears. AI training can therefore not only teach technical skills, but also change the corporate culture to enable the sustainable use of AI.

Keywords: Artificial Intelligence (AI), production, employee training, corporate culture

Introduction

Artificial intelligence (AI) is currently being discussed as the next ‘general purpose technology,’ i.e., a technology with a wide range of applications. Manufacturing companies see AI as the decisive factor in their competitiveness. In addition to technological challenges, there are also organisational challenges to the meaningful use of AI in manufacturing companies, as these technologies must be understood and mastered by employees. For many, the biggest hurdle is not technical, but organisational and cultural. Training courses focusing on technological aspects are an important building block in enabling employees to use AI at an early stage, thereby increasing acceptance of AI applications [1]. Other factors that influence the beneficial use of AI applications are the digital maturity of the company and the individual willingness of employees to embrace digitalisation [2], [3], which are also reflected in the

corporate culture [1], [4]. This article aims to show how training can help create the conditions for the long-term use of AI applications in manufacturing companies.

Looking at existing AI training concepts, these primarily focus on technical aspects of AI application. The aim is to provide training participants with a basic understanding of AI. Basic technical knowledge and various terms and definitions are taught (what is AI, how do machine learning and deep learning differ, what can AI do – and what can't it do?) in order to be able to classify AI appropriately. Application examples (e.g. from production) are often used to promote understanding by linking basic knowledge with an application. Technical basics also include data literacy [5], i.e. the ability to understand data (through observation, analysis, evaluation, reflection) and use it (integration, navigation, collection, presentation) [5]. With regard to AI applications, employees must understand how data must be prepared for AI use (e.g. for predictive maintenance, quality assurance, automation) and, above all, what results this data can deliver [5]. However, data literacy also includes the ability to handle data responsibly. In manufacturing companies, employees must be able to apply this data literacy to processes in which data is used.

In order to identify the potential of AI, it is also necessary to understand the processes that contribute to the company's value creation [6]. These processes are aligned with corporate goals such as efficiency and effectiveness, but also customer orientation, yet they are rarely addressed directly in training concepts. By using practical use cases, process thinking can also be taught in addition to the technical aspects. In this way, the changes in roles and processes resulting from the use of AI at the interface between people, processes and technology can also be taught. Training can also be used to alleviate fears, for example by involving role models (e.g. IT-savvy employees, data analysts, maintenance staff). If real data (e.g. from the company itself) is used in the training courses, this also increases confidence in handling the company's own data and leads to initial successes. This increases motivation and promotes acceptance.

Another prerequisite for successful AI implementation is an appropriate corporate culture [1], [4]. Managers play a particularly important role in this context. They actively contribute to eliminating uncertainties and enabling change to take place. The uncertainties that need to be dispelled primarily concern the possible loss of jobs due to AI, but also the possible loss of control over one's own tasks. The corporate culture should also allow employees to be actively involved in the company's AI strategy and to participate in shaping AI projects. If this involvement takes place at an early stage, acceptance also increases [1]. The corporate culture should also promote integrative approaches that involve different people with different levels of knowledge, as these contribute to raising awareness of the processes and interrelationships [1], [4]. In manufacturing companies in particular, where departments often work within fixed structures, AI requires cross-departmental and cross-functional collaboration in order to exploit potential. Training can therefore not only increase the technical knowledge of employees, but also eliminate uncertainties, provided that the appropriate corporate culture is in place [4].

Procedure

This article is based on aggregated data from companies that have implemented (at least) one TBI as part of EDIH AI5production and undertaken corresponding training measures. The basis is provided by 151 TBIs (as of summer 2025) that were subjected to a standardised measurement of digital maturity (Digital Maturity Assessment – DMA) using a tool (Digital Maturity Assessment Tool) as part of the project [7], [8]. This records the current state of digital maturity at a specific point in time in order to determine in which areas the company might need support. The measurement is carried out at three points in time – before the TBI (DMA-T0), one year after the TBI (DMA-T1) and at a later point in time (DMA-T2). Even without the involvement of an EDIH, the OpenDMA tool can be used by companies to clarify their own digital maturity [7]. Six dimensions are recorded, which are subdivided into further sub-dimensions. Measurements are taken using questionnaires that can be either binary (yes/no) or Likert-scaled (1–5). The following six dimensions are recorded:

- (1) The company's digitalisation strategy – in relation to planned or actual investments
- (2) Readiness for digital transformation – assessment of the current use of digital technologies, including mainstream and more advanced technologies
- (3) The company's digitalisation strategy – in relation to planned or completed investments
- (4) Readiness for digital transformation – assessment of the current use of digital technologies, including mainstream and more advanced technologies
- (5) People-centred digitalisation – based on how employees are trained, involved and empowered for and through digital technologies, including improving working conditions with the aim of increasing their productivity and well-being
- (6) Data management and connectivity – type and organisation of digital data storage within the company and its use for business purposes, taking into account data protection and cyber security
- (7) Automation and artificial intelligence – examines the extent to which digital means enable automation and knowledge gain in business processes
- (8) Green digitalisation – captures the ability to implement digitalisation with a long-term concept, taking sustainability aspects into account

The degree of achievement for each dimension is shown as a percentage, and a DMA score is calculated across all dimensions. More than half of the TBIs recorded have already undergone two measurements (DMA-T1). This article focuses primarily on 46 TBIs in which the possibilities for AI use were tested. In addition, training measures undertaken by the companies were documented and analysed as part of the EDIH AI5production programme.

Results

The average digital maturity score (DMAScore) was 51.4% in the first measurement and rose to 59.5% after a TBI (DMA-T1) was carried out. If only companies with two measurements (DMA-T0 and DMA-T1) are included, the average DMA rose by 9.3% and in the individual dimensions by 7.9% (company digitalisation strategy), 8.0% (readiness for digital change), 12.0% (people-oriented digitalisation), 9.3% (data management and connectivity), 10.1% (automation and artificial intelligence) and 8.9% (green digitalisation). Figure 1 provides an overview of the changes in the DMA of companies with two measurements (DMA-T0 and DMA-T1). Focusing on the TBIs that dealt with the possible applications of AI, these are above average in all dimensions. In the dimension 'Readiness for digital change', the result rose by 9% and in the dimension 'Data management and connectivity' by 11%. The highest increase, at 15%, was in the dimension 'Human-centred digitalisation'. This dimension covered specific training plans for retraining and/or further training of employees offered by the companies. In addition, planned measures and offers – from regular, comprehensive training courses on advanced technologies or the acquisition of digital Skills – identified. These are tailored to the specific needs and training requirements of employees. Companies that have implemented these measures have also seen improvements in other areas. Interactive training courses that offer extensive opportunities to ask questions and include practical examples are particularly promising. The EDIH AI5production also made it clear that workshops with interdisciplinary teams from manufacturing, IT and management increase the acceptance of new technology (especially AI), as fears are reduced and concrete use cases are developed together. Working together also creates a stronger understanding for and among each other. This requires an open corporate culture, which is reflected, for example, in a flat hierarchy, support from management and active integrative approaches (e.g. in the form of cross-functional groups). In addition, an active error culture, the identification of role models and clearly defined corporate principles for the use of AI are also important factors.

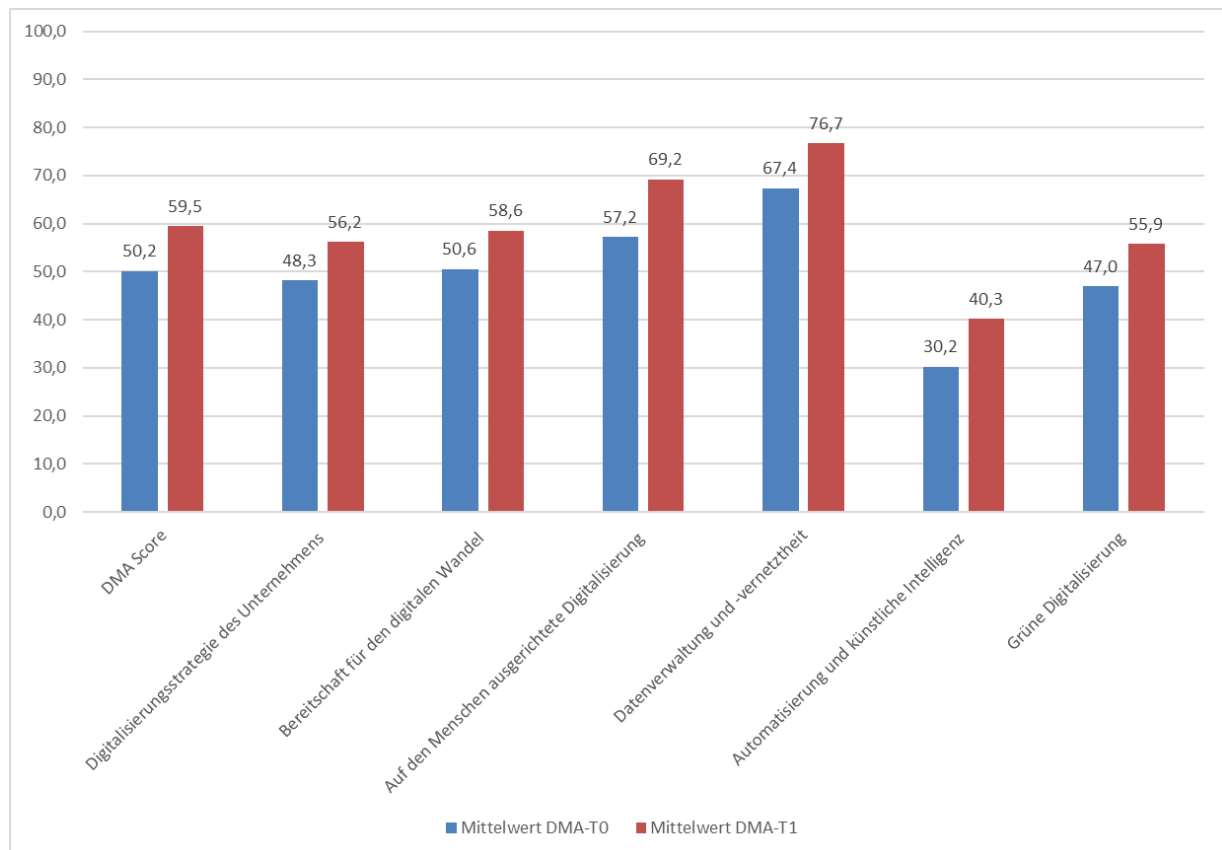


Figure 1. Changes in DMA Score and dimensions of companies with at least 2 measured values

Conclusion and recommendations for action

In order to introduce AI into manufacturing companies and achieve corporate goals such as increased efficiency and effectiveness, as well as customer satisfaction, it is necessary to familiarise employees with this new technology at an early stage. Training courses usually cover the technical aspects in particular, but by taking organisational aspects into account, a holistic change can be enabled. In addition to the technical basics of AI, aspects such as data literacy and security must be addressed in the training courses. Participants should understand that they are responsible for the data they use in business processes. Corporate culture plays a particularly important role in traditional companies where cross-departmental work was not absolutely necessary. Corporate culture also contributes to the respectful interaction that is necessary for collaboration in the context of AI. In order to create a common understanding, training courses must be designed and coordinated accordingly. The results from EDIH AI5production have shown that practical, interactive and integrative workshops can play an important role here. Managers can act as multipliers to reduce fears and increase acceptance of the technology. Sie sind auch dafür verantwortlich als Teil der Unternehmenskultur ein Klima zu schaffen, in dem Fragen und Zweifel ernst genommen werden. You are also responsible for creating a climate in which questions and doubts are taken seriously as part of the corporate culture.

For companies that want to use AI, recommendations for action can be derived from the results of the EDIH AI5production. When getting started with AI, it is advisable to first analyse the current situation in the company, for example as part of a TBI with integrated training. To this end, it makes sense to bring together mixed teams from all areas (production, IT, technology, management). In this way, AI is not understood as an isolated technology, but as an integral part of the value chain. In addition, it is advisable to develop 'AI ambassadors' who serve as contact persons – an approach that has led to a noticeable improvement in the learning and innovation culture in TBIs. The introduction of AI is always directly related to the corporate culture, which must be characterised by participation and openness.

At the end of the training programme, a team may emerge that not only understands how to use AI, but also critically questions and actively shapes its application. The strong practical orientation and focus on specific use cases make this concept particularly suitable for SMEs that want to introduce AI in a meaningful and resource-efficient way. The goal must be for the corporate culture to contribute to successful AI implementation by creating a positive attitude towards change and increasing acceptance.

References

- [1] N. Rane, S. P. Choudhary, und J. Rane, „Acceptance of artificial intelligence: key factors, challenges, and implementation strategies“, *Journal of Applied Artificial Intelligence*, Bd. 5, Nr. 2, S. 50–70, Sep. 2024, doi: 10.48185/jaai.v5i2.1017.
- [2] C. Williams, B. Krumay, D. Schallmo, und E. Scornavacca, „Digital maturity model for SMEs: validation through a mixed-method approach“, *Pacific Asia Journal of the Association for Information Systems*, Bd. 16, Nr. 1, S. 2, 2024.
- [3] M. Muehlburger, B. Krumay, S. Koch, und S. Curre, „INDIVIDUAL DIGITAL TRANSFORMATION READINESS: CONCEPTUALISATION AND SCALE DEVELOPMENT“, *Int. J. Innov. Mgt.*, Bd. 26, Nr. 03, S. 2240013, Apr. 2022, doi: 10.1142/S1363919622400138.
- [4] O. T. Murire, „Artificial Intelligence and Its Role in Shaping Organizational Work Practices and Culture“, *Administrative Sciences*, Bd. 14, Nr. 12, S. 316, Dez. 2024, doi: 10.3390/admsci14120316.
- [5] L. Van Audenhove, L. Vermeire, W. Van den Broeck, und A. Demeulenaere, „Data literacy in the new EU DigComp 2.2 framework how DigComp defines competences on artificial intelligence, internet of things and data“, *Information and Learning Sciences*, Bd. 125, Nr. 5–6, S. 406–436, Feb. 2024, doi: 10.1108/ILS-06-2023-0072.
- [6] M. Muehlburger und B. Krumay, „Towards a context-sensitive conceptualisation of digital transformation“, *Journal of Information Technology*, Bd. 39, Nov. 2023, doi: 10.1177/02683962231219514.
- [7] „Open DMAT - Digital Maturity Assessment tool | European Digital Innovation Hubs Network“. Zugegriffen: 5. Oktober 2025. [Online]. Verfügbar unter: <https://european-digital-innovation-hubs.ec.europa.eu/open-dma>
- [8] F. Sümeç, C. Wögerer, und G. Hesina, „Der Weg zur digitalen Transformation in der Produktion, EDIH (European Digital Innovation Hubs Network)“, *Technische Universität Wien, Pilotfabrik Industrie 4.0*, 2025.

Benefits for manufacturing companies through the European innovation ecosystem

Stefanie Werderits
Plattform Industrie 4.0 Österreich
stefanie.werderits@plattformindustrie.at

Roland Sommer
Plattform Industrie 4.0 Österreich
roland.sommer@plattformindustrie40.at

Kurzzusammenfassung

Der European Digital Innovation Hub (EDIH) AI5production bietet ein starkes, kooperatives Umfeld für KMUs und Mid-Caps – regional, national und europäisch – zur Förderung von Digitalisierung und Innovation. Strategische Auswahl von Partnerschaften mit anderen EDIHs und Stakeholdern, organisationübergreifende sowie grenzüberschreitende Wissensvermittlung und gemeinsame Aktivitäten wie Webinare oder themenbezogene Netzwerk-Events ermöglichen den Austausch von Erfahrungen und Praxisbeispiele (Best Practices). Dadurch wird auch die Servicequalität der EDIHs verbessert. Die Erfahrungen, Herausforderungen und Potenziale im Aufbau von Netzwerken im Europäischen Innovationsökosystem aus dem Arbeitspaket „Innovationsökosystem und Netzwerken“ des EDIH AI5production werden in diesem Beitrag dargestellt, mit besonderem Fokus auf Mehrwert, Synergien und konkrete Unterstützungsangebote für Unternehmen.

Keywords: Netzwerk, EDIH, Innovationsökosystem, Kooperation, Digitalisierung

Abstract

The European Digital Innovation Hub (EDIH) AI5production offers a strong, collaborative environment for SMEs and mid-caps – regionally, nationally and across Europe – to promote digitalisation and innovation. Strategic partnerships with other EDIHs and stakeholders, cross-organisational and cross-border knowledge transfer, and joint activities such as webinars or themed networking events enable the exchange of experiences and best practices. This also improves the service quality of the EDIHs. This article presents the experiences, challenges and potential in building networks in the European innovation ecosystem from the ‘Innovation Ecosystem and Networks’ work package of the EDIH AI5production, with a particular focus on added value, synergies and concrete support services for companies.

Keywords: Networks, EDIH, innovation ecosystem, cooperation, digitalization

Introduction

Digital transformation and innovation cannot be achieved alone – they require functioning networks [1], [2]. However, many companies find it difficult to access tailored technical and strategic support in the digitalisation process. They often lack information about technological potential, available funding opportunities, suitable partners or practical examples. One of the goals of EDIH AI5production is to develop and expand a strong national and European innovation ecosystem for SMEs and mid-caps. Embedding them in a functioning innovation ecosystem and existing and new networks can help accelerate the innovation developments of SMEs and compensate for resources and knowledge [3]. Innovation ecosystems can be defined as systems with heterogeneous but complementary actors [4]. These are not organised hierarchically, but create value through modular structures and loosely coupled collaborations. The EDIH AI5production acts as an orchestrating actor in the sense of an ‘ecosystem orchestrator’ – i.e. an entity that establishes framework conditions and interfaces [4]. This role underlines the importance of complementarity, coordination mechanisms and targeted partner integration in

digital innovation networks [4]. Innovation in ecosystems arises from loosely coupled but coordinated actors [4] who are connected to each other through clearly defined interfaces and offer complementary services (e.g. consulting, technical services and financing). Networks thus create value through access to knowledge, interfaces and cooperation [4]. The challenges of collaboration in an ecosystem include issues such as demarcation, overlapping roles, access to resources and trust [4]. The aim is to implement the elements of a successful ecosystem within the framework of AI5production for the benefit of SMEs and mid-caps.

Procedure

Based on the previously presented fundamentals of ecosystems, the role of EDIH AI5production (in particular the work package ‘Innovation Ecosystem and Networks’) was identified and used for the further development of the ecosystem. First, the initial situation of AI5production is presented, followed by a conceptual classification of the role of AI5production.

Initial situation

The EDIH network, consisting of SMEs, research, politics and intermediaries, exhibits both modularity and complementarity. The network consists of diverse, often decentralised actors who complement each other with their complementary skills. The EDIH AI5production consortium also consists of partners who complement each other with their wide-ranging expertise, thus fulfilling the requirements of an ‘innovation ecosystem’. In addition, modularity and complementarity are also taken into account in the individualised design of the collaboration and in the targeted selection of partners for EDIH collaborations. A specific challenge in the AI5production project to date has been access to resources due to the different national funding structures of the EDIHs (e.g. national requirements for resource use limited to SMEs in the respective nation state) and language barriers among SMEs.

Documentation of the approach

The core of the strategy and central task of EDIH AI5production is to promote and expand European networks with the aim of offering SMEs and mid-caps tailored support through access to the Austrian and European innovation ecosystem and corresponding networking opportunities (see Figure 1). The networking strategy comprises the following activities:

- Targeted selection of partners according to topic and industry relevance
- Establishment of personal contacts and formal cooperation via memoranda of understanding (MoUs)
- Development of joint activities (e.g. webinars, dissemination events)
- Active communication of relevant information and services via the network

The establishment and expansion of the ecosystem and network services is hybrid in design, combining the strategic planning behind the network development with dynamic adaptation to technology trends and needs. These needs have also been repeatedly expressed and implemented by the companies. A clear example of this is the demand for information on the possible use of large language models (LLMs) in production. This need was expressed and taken up by a number of companies in the course of the activities. The activities are documented, evaluated and continuously adapted to new needs at SMEs.

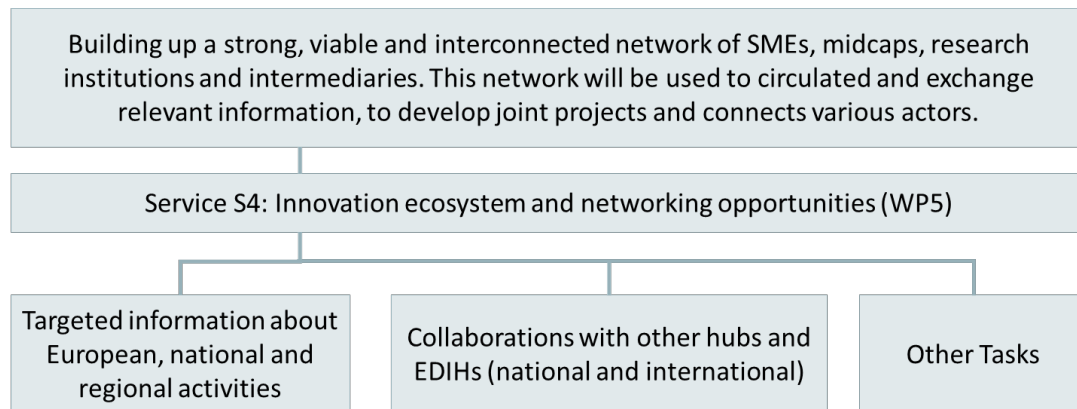


Figure 1. Objectives in the innovation ecosystem and networks work package

The strategic selection of partnerships is based on certain factors such as production relevance, but also thematic and geographical proximity. At the same time, partnerships that have formed organically during the course of the project also contribute to strengthening the network. The activities within the framework of the collaborations are designed individually and tailored to the respective expertise and framework conditions of the individual EDIHs. Partnerships with stakeholders outside the consortium are defined via MoUs, but are not hierarchically controlled. Access to knowledge, interfaces and cooperation is provided through activities such as webinars, project initiatives or the mediation of suitable cooperation partners. This ‘value creation through connection’ is a key objective of EDIH AI5production. Integration into the European ecosystem serves to disseminate relevant information, exchange best practices, develop joint projects and network various stakeholders (including Austrian SMEs and mid-caps).

Results

There are a number of results from the AI5production project, in which a functioning cooperation network was established in Austria and Europe.

Regional and national network development

At the regional level, the topic-related training formats in attendance with the participation of various network partners are an important starting point for companies to take advantage of the services of EDIH AI5production in a next step. These formats, which go beyond pure dissemination events (e.g. trade fair stands, presentations about the project), are innovative in nature, as they make targeted use of the potential of AI5production and its partners and combine successful event elements. On the one hand, the broad partner network and its contacts make it possible to bring researchers together with industry representatives and, at the same time, to address regions of Austria that are often characterised by a low density of events. By specifically activating the partners' networks and involving relevant multipliers, such as interest groups, it is possible to attract the attention of SMEs and mid-caps. On the other hand, the combination of thematically focused information with the opportunity for direct exchange with experts and like-minded people has met with a positive response. In order to appeal to local SMEs, these events are deliberately held on site in different regions of Austria.

The basic idea behind this event format is to present and communicate current technologies in an understandable way. Researchers take on this task in the format, while industry representatives use concrete examples (‘use cases’) to demonstrate real-world industrial applications. The representatives of the hubs (EDIH, DIH) act as ‘honest brokers’, which means – as is clearly communicated at the beginning of the events – that the focus is not on product presentations or any kind of commercialisation, but rather on providing reliable, ‘neutral’ information. The events are accompanied by a presentation of further Austrian funding programmes, a showcase of pilot factories and research centres on these specific topics, and a list of contact persons. A central element is the facilitation of intensive discussions

between the participants and the speakers following the presentation part. Last but not least, experience gained from previous TBIs helps to identify current topics for the events. The quality of the events can therefore be attributed to the unique network of partners and its targeted expansion, as well as to direct feedback from participants and the seamless transition to other services. Topics addressed included industrial data exchange, the digital product passport, generative artificial intelligence and operational energy management.

Another focus was the targeted involvement of various industrial networks in Austria, for example in mechanical engineering and the electrical and electronics industry. With these players, a continuous transfer of knowledge between EDIH AI5production and its network partners was established. For better orientation in the innovation ecosystem, AI5production presents relevant flagship projects, regulations and further initiatives on its website. The aim behind this is to prepare results from nationally or European-funded flagship projects in a way that is suitable for SMEs. In many cases, flagship projects deal with specific use cases whose results may be relevant for other companies. The topics addressed in these projects are also dealt with in depth at events and training courses. Thanks to the support of the network partners, regional training courses can be offered at regular intervals, for example, which deal with current regulations and technologies in selected theoretical and practical contributions.

The mutual involvement of network partners in sending out and promoting events, as well as in presenting use cases, raises awareness of the EDIHs' offerings among SMEs and mid-cap companies throughout Austria. The 'inform, participate and try out' approach promoted by this component of the cooperation network is therefore a success factor in reaching SMEs and connecting them with partners.

The European network – increasing cooperation between EDIHs

The EDIH AI5production has established a number of European partnerships to improve access to digital technologies and application knowledge and leverage synergies between EDIHs through the exchange of experience, events and targeted training. The establishment of the European network was supported by memoranda of understanding. There was an intensive exchange between the EDIHs on the various services they offer. This was a key activity, particularly in the early stages of the EDIH, in order to learn from each other and synchronise activities where necessary. In addition to the exchange of services, the focus was also on other issues, such as the target groups to be addressed, billing modalities, reporting requirements and experiences with interim evaluations.

Another focus was on providing SMEs and mid-caps with access to industrial best practices and topic-specific expertise, including across national borders. Webinars organised in collaboration with one or more EDIHs have proven to be particularly effective and practical in implementation. The advantage of these transnational webinars is that they provide companies with low-threshold access to expert knowledge and inspiring practical examples, enabling them to benefit from the specific strengths and experiences of different hubs. These formats offer SMEs and mid-caps European know-how on technologies, best business practices and successful use cases (e.g. successfully implemented TBIs) that addressed current topics, including digital twins, security and artificial intelligence.

The design of the first webinars with European partner EDIHs was based on the joint identification of relevant topics and was developed in the course of individual discussions with the EDIHs. The idea of involving several hubs in a thematic webinar in order to gain insights into different best practices and access to experts also arose from an exchange of ideas between the EDIHs. This demonstrates the potential for collaborative innovation among EDIHs. A physical counterpart and example of success is the Czech-Austrian EDIH Twister in Brno [5]. In addition, the event format of an 'EDIH workshop', where EDIHs met on site in Vienna, provided an opportunity to generate ideas for joint services. In order to facilitate low-threshold exchange between EDIHs in a similar way and to benefit from their insights into the needs of SMEs and mid-caps, the idea of topic-specific online workshops with partner EDIHs was launched. A webinar and workshop on the topic of Digital Product Passport (DPP) is currently being planned. This will be organised and implemented in cooperation with the EU project Cirpass-2 [6]. A small-scale pilot will test whether this format can contribute to the roll-out of the DPP

and to what extent the EDIHs can play a role in supporting SMEs and mid-caps in the roll-out of the DPP. A targeted reflection on the role and possibilities of the EDIHs in supporting the implementation of technologies and regulations can therefore offer additional assistance to companies throughout Europe.

Cooperation between EDIHs also promotes the exchange of information on current developments, such as the needs of businesses, information on other projects, regulations and initiatives that may be relevant to SMEs and mid-cap companies in the EU. The EDIH network is well suited for the effective dissemination of information and offers, for example, the opportunity to communicate the recently launched 'SECURE Project' [7] for the implementation of the 'Cyber Resilience Act' to the target group.

Often, the budget of EDIH cooperation partners is reserved exclusively or predominantly for national companies (based on experience from EDIH AI5production), which limits the possibilities for mutual support – i.e. not all EDIHs are allowed to make their services available to SMEs and mid-caps across the EU. Nevertheless, the exchange of best practices, for example on the activities and design of EDIH services or on successful practical examples, provides valuable insights that contribute to improving the quality of EDIH AI5production.

Cooperation with other EDIHs can offer added value in terms of access to certain services, particularly training and TBIs, but also in the area of financing – provided that this is possible under national funding agreements. Cooperation between EDIHs enables SMEs and mid-caps to connect with a broad network via a local hub, gain access to the expertise of the partners represented in it, and potentially benefit from services offered beyond Austria's borders. In this way, the most suitable support from a partner with specific expertise can be identified within the network. However, language barriers and geographical distances should not be ignored as obstacles. In practice, SMEs and mid-caps have therefore only been able to make limited use of certain services offered by other European hubs.

Supraregional network

In addition to establishing a Europe-wide ecosystem, another focus is on creating a seamless network in Austria's border regions and on supraregional cooperation. There is cooperation with EDIHs in Slovakia, Slovenia, the Czech Republic, Germany and Hungary. Cooperation is also being established with two EDIHs in Ukraine. Within Austria, EDIH AI5production is well connected with the other Austrian EDIHs and the national DIHs (Digital Innovation Hubs).

Conclusion

Companies are facing major structural challenges resulting from ever faster innovation cycles, the disruptive elements of new technologies in the production process, especially artificial intelligence, the adaptation of organisational structures to rapid technology cycles and automation requirements, the need for further training and higher qualifications for employees, and the need to implement data-based business models. Industry 5.0 is essentially concerned with a human-centred approach and thus with organisational structures and training. The focus on aspects of Industry 5.0 is deliberately at the heart of EDIH AI5production – as expressed by the number '5' in the title.

The services offered by EDIH provide companies with the opportunity to implement comprehensive, individualised and, at the same time, low-threshold structural measures for digital transformation. This sets the format apart from other support measures. The innovation of the solution lies in the fact that different services are individually tailored to the needs and requirements of the companies. EDIH AI5production has set up its internal processes in such a way that, for example, companies receive feedback on TBI applications within a maximum of 72 hours and can therefore start their activities very quickly.

Expected benefits for companies

Participating companies benefit from better orientation in the European innovation environment, access to information, training and Europe-wide offers, a focused network of contacts and the resulting project

opportunities. The activities in the ‘Innovation Ecosystem and Networks’ work package play a key role in helping SMEs and mid-caps recognise the benefits of EDIH services (TBI, training opportunities and STFI), thereby contributing to the digital transformation of companies. These activities also play a key role in enabling EDIH AI5production to initially reach potentially interested SMEs and mid-caps by specifically involving existing networks in its activities. Through formats that promote exchange between companies of different sizes (e.g. Large companies and SMEs) along the supply chain (including suppliers) and experts from various institutions (science, consulting, etc.) can promote networking within an industry. Beyond EDIH AI5production, the contacts made can benefit the companies. In the future, the aim is to further expand regional activities and international partnerships, deepen exchanges and test new forms of cooperation. A stronger focus on artificial intelligence also addresses a development that is important for all companies. The aim is to contribute to a higher level of innovation and digitalisation with every activity and thus achieve added value for SMEs and mid-cap companies in Europe – which is expected to have spillover effects along the value chains.

The EDIH AI5production network shows that when European cooperation is seen not as an obligation but as a creative opportunity, a resilient innovation ecosystem emerges. The next steps in enhancing cooperation consist of expanding thematic focus activities (e.g. AI, circular economy through digitalisation, cyber security), identifying practical examples and cross-EDIH offerings, providing an overview of existing offerings, and specifically addressing current challenges. In addition, deepening regional networks and specifically targeting local SMEs and mid-caps is a key success factor for a resilient innovation ecosystem.

References

- [1] A. Bargoni, A. Ferraris, Š. Vilamová, und W. M. H. Wan Hussain, „Digitalisation and Internationalisation in SMEs: a Systematic Review and Research Agenda“, *Journal of Enterprise Information Management*, Bd. 37, Nr. 5, S. 1418–1457, 2024.
- [2] „European Innovation Ecosystems - Research and innovation“. Zugegriffen: 7. August 2025. [Online]. Verfügbar unter: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/european-innovation-ecosystems_en
- [3] F. P. Hilmersson und M. Hilmersson, „Networking to accelerate the pace of SME innovations“, *Journal of Innovation & Knowledge*, Bd. 6, Nr. 1, S. 43–49, 2021.
- [4] M. G. Jacobides, C. Cennamo, und A. Gawer, „Towards a theory of ecosystems“, *Strategic Management Journal*, Bd. 39, Nr. 8, S. 2255–2276, Aug. 2018, doi: 10.1002/smj.2904.
- [5] „AI5production Erfolgsstories - Networking“, EDIH AI5production. Zugegriffen: 7. August 2025. [Online]. Verfügbar unter: <https://ai5production.at/success-story/network/>
- [6] Plattform Industrie 4.0, „Exploring the Role of EDIHs in Supporting SMEs with the DPP“. Zugegriffen: 30. September 2025. [Online]. Verfügbar unter: <https://plattformindustrie40.at/blog/event/exploring-the-role-of-edih-in-supporting-smes-with-the-dpp/>
- [7] Secure, „SECURE Strengthening EU SMEs Cyber Resilience“, SECURE. Zugegriffen: 30. September 2025. [Online]. Verfügbar unter: <https://secure4sme.eu/>

Artificial intelligence in production

Franz Sümecz
Pilotfabrik, Technische Universität Wien
franz.suemecz@tuwien.ac.at

Barbara Krumay
JKU Business School, Johannes Kepler
Universität Linz
barbara.krumay@jku.at

Kurzzusammenfassung

Künstliche Intelligenz spielt in der Produktion eine immer größere Rolle. Das zeigt sich auch in den TBIs, die im Rahmen des EDIH AI5production umgesetzt wurden. Allerdings stehen Unternehmen vor der Herausforderung aus dem immer größer werdenden Angebot die KI-Anwendung auszuwählen, die die aktuelle Problemstellung auch lösen kann. Wissenschaftliche Kategorisierungen können helfen, um diese Auswahl zu vereinfachen. Zusätzlich können Beispiele (Use Cases) genutzt werden, die zusätzlich zu diesen Kategorisierungen als Orientierung für ähnliche Problemstellungen dienen können. Insbesondere Anwendungen, die unterschiedliche Technologien integrieren können aus diesen Beispielen lernen. Im Rahmen dieses Beitrags soll gezeigt werden, wie Kategorien und Use Cases für Unternehmen die Selektion von entsprechenden Ansätzen unterstützt werden kann.

Keywords: Künstliche Intelligenz, Produktion, Anwendungen, Kategorisierung

Abstract

Artificial intelligence is playing an increasingly important role in production. This is also evident in the TBIs that were implemented as part of EDIH AI5production. However, companies are faced with the challenge of selecting the AI application that can solve their current problem from an ever-growing range of options. Scientific categorisations can help to simplify this selection process. In addition, examples (use cases) can be used to provide guidance for similar problems in addition to these categorisations. Applications that can integrate different technologies can learn from these examples in particular. This article aims to show how categories and use cases can support companies in selecting appropriate approaches.

Keywords: Artificial intelligence, production, application, categorization

Introduction

The importance of systems based on artificial intelligence (AI) continues to grow. They will be indispensable for the production of the future. Back in 2018, three out of four companies expected AI applications to revolutionise their core business within the next three years [1]. Forecasts predict that generative AI in particular could increase global gross domestic product by 7%, but at the same time 300 million knowledge worker jobs could be lost [2]. Companies expect that the use of AI will increase efficiency and effectiveness [3]. However, it has become apparent that selecting a suitable AI-based system is difficult, as the desired results and long-term benefits are hard to assess. Large and disruptive AI projects often fail, while smaller, application-oriented projects (known as ‘low-hanging fruit’) based on existing applications are more promising [1], [4]. The multitude of possible applications – from analytical tasks and real-time decision support to automation in the production hall – means that the choice goes beyond the conventional make-or-buy decision [5]. The focus for this choice is usually a real-world problem. This problem cannot usually be solved by one technology alone, but only through the interaction of different technologies (e.g. the Internet of Things and AI in the context of Industry 4.0) [5]. It is irrelevant for this decision whether the term AI refers to machine learning (ML), deep learning or generative AI. What is important is to understand that AI-based systems are not an end in themselves, but can only be used effectively if the underlying problem and the goal to be achieved with the use of AI-based systems are clear. These objectives include maximising operating results, increasing sales, reducing costs, strengthening competitiveness, improving resilience, particularly in global value chains, and strengthening sustainability (resource conservation, circular economy, energy transition) [1], [3],

[5], [6]. The introduction and subsequent use of AI-based systems – especially for manufacturing companies – is based, among other things, on the possibility of testing the systems and their applicability on a small scale [1], [7]. The TBIs carried out in EDIH AI5production are a suitable approach for this. Based on the problems addressed in the TBIs, categories were identified that can be used as a basis for decision-making by other manufacturing companies.

Artificial intelligence

The European Union defines an AI system as "a machine-based system designed to operate with varying degrees of autonomy and demonstrate adaptability after deployment, which derives explicit or implicit objectives from the inputs it receives, such as outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments" [8]. AI imitated human cognitive abilities (intelligence) to automatically solve complex problems that previously could only be solved through human intervention (human intelligence) in the form of assessment and decision-making [9]. The first scientific sources on the subject of AI date back to 1955 [10] – since then, our understanding has not changed fundamentally, but the possible applications have. The basis for all AI-based systems is data, which can be in structured or unstructured form. As with all data-based applications, the quality of the data influences the results of AI-based systems ('garbage in, garbage out'). Various methods (e.g. statistical methods) are applied to this data to solve a specific problem. A fundamental difference to data processing without AI is that AI is "trained" on a small portion of the data in order to develop ways of interpreting the entire database [5]. Existing categorisations of AI-based systems mainly refer to the way in which data is used. The relationships are often represented in a multi-layered model. AI forms the outer layer with machine learning (ML) as a subset of AI and deep learning (DL) as a subset of ML.

So-called generative AI has taken on particular significance. Generative refers to the automated generation of new texts, images, videos or music [6]. In this context, 'new' means that the result does not exist in this form in the underlying data. Many applications of generative AI are based on direct interaction with humans. This interaction is usually enabled in the form of a prompt, i.e. the desired result is described by the human in natural language and in text form in a text-based interface. From this, a result is automatically generated that corresponds as closely as possible to this description. The result can be further refined in a dialogue (chat) between humans and AI. Depending on the model used, the result can be specifically a text (e.g. ChatGPT - GPT-o1) or an image (text-to-image generators such as Midjourney, DALL-E, Stable Diffusion). Some AI-based systems also focus specifically on generating programme code and thus take on software development tasks (e.g. Cursor). Over time, the offerings have evolved and become more generally applicable, meaning they can deliver different results (e.g. ChatGPT 5.0, Google Gemini). The integration of AI-based systems into applications is also constantly evolving. This includes systems integrated into office applications (e.g. Microsoft's Co-Pilot, AI-supported Adobe Acrobat functions for text creation and editing) as well as AI as part of control systems [5]. Generative AI makes everyday office life much easier, especially when it comes to creating a wide variety of documents. This is currently the best-known form of AI application, which has become particularly widespread in the consumer sector. Large language models (LLMs) have become established here, i.e. systems that can generate new text from large, text-based data using statistical methods. Kategorisierung von KI-basierten Systemen

In order to increase understanding of AI-based systems and to enable structured selection, various categorisations have been made. Many category schemes are based on which methods are applied to the data. One approach that has emerged in the context of manufacturing companies focuses on machine learning (ML) as the top level and further categorises according to learning paradigms, learning models and learning tasks, assigning applications in manufacturing companies to these categories (see Figure 1) [5].

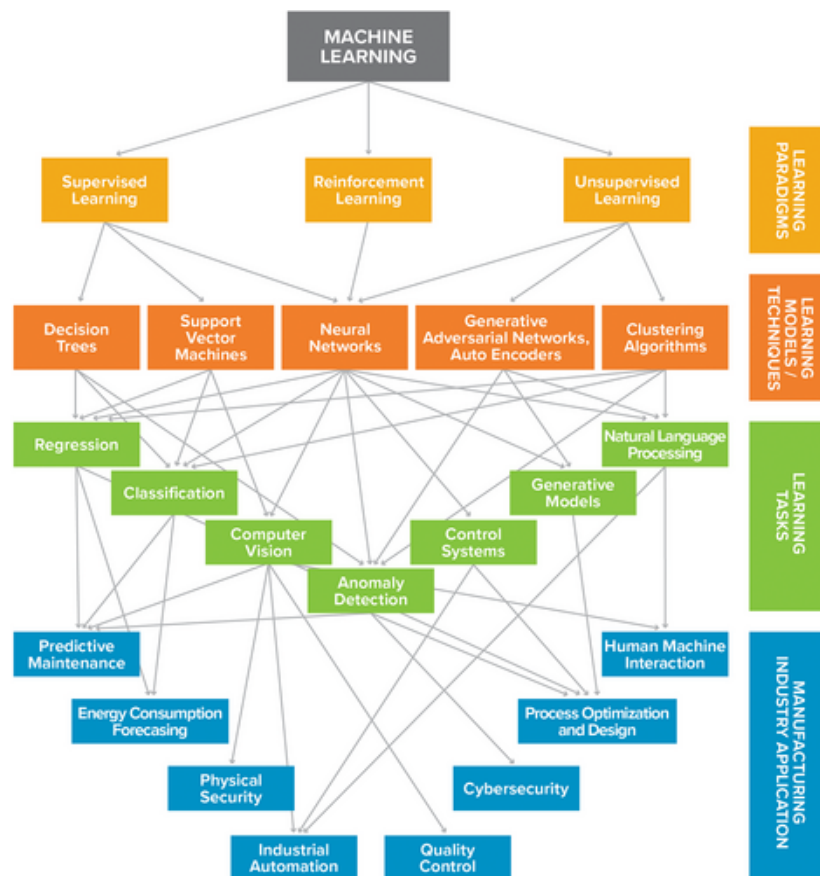


Figure 1. Kategorisierung von ML inkl. Anwendungen in der Produktion [4]

Another categorisation, which refers to problems in companies in general, distinguishes only three categories [1]: (1) automating business processes, (2) gaining insight through data analysis, and (3) engaging with customers and employees [1]. Based on rule-based problem solving without AI (such as business process automation), AI enables a different form of problem solving that more closely resembles the human approach. AI support extends the automation of business processes towards robotic process automation (RPA), which can, for example, read and interpret contracts in order to initiate process steps. The same applies to AI-supported data analysis, which, primarily with the help of machine learning approaches, can recognise certain patterns ('pattern recognition') and interpret them in order to predict necessary maintenance work ('predictive maintenance'), for example. The last category describes text-based applications (e.g. chatbots) that can be used in customer service, for medication recommendations for patients, and also in internal knowledge management.

Implementation of AI-based systems

As already mentioned, the implementation and long-term use of AI-based systems is not always successful. Experience has shown that this is particularly successful when AI-based systems are implemented as part of or in addition to existing applications [1]. Four key factors have been identified: (1) understanding of technology, (2) project portfolios, (3) piloting, and (4) scaling [1]. Understanding technology means that the company must have a certain level of understanding of AI technologies, especially with regard to their strengths and weaknesses. Since, unlike rule-based systems, complete transparency cannot be achieved, trust is also an important issue. In addition, a project portfolio must be in place that covers the possibilities of AI-based systems in the company and demonstrates them using use cases. Piloting plays a special role here, i.e. AI-based systems must be tested on a small scale in order to assess their effectiveness in terms of proof of concept, taking regulations into account, before large investments are made. This also strengthens confidence in the AI-based system. If the AI-based system is actually implemented after the pilot phase, care must be taken to ensure that it is scalable, i.e.

that it can be used company-wide and is robust and expandable for future changes (e.g. due to more data) [1].

Procedure and Results

The EDIH AI5production project also aims to demonstrate how digitalisation (in the broadest sense of the term) can be used to increase production efficiency. In other words, the aim is to show how the use of digital technologies can contribute to achieving the stated corporate goals. Over 150 TBIs have been carried out within the framework of EDIH AI5production in three years. This clearly shows that AI methods have long since become an integral part of digitalised production processes and production workflows. 45% of the TBIs carried out in EDIH AI5production test technical solutions using AI (hereinafter referred to as TBIs with AI application). These TBIs were analysed based on existing category schemes [1], [5]. TBIs that contain characteristics of several categories were assigned multiple times. In a further analysis, categories were specified with the help of selected TBIs. An overview can be found in Table 1 (multiple assignment). Almost half of TBIs with AI applications use image-based AI, and around 30% of TBIs with AI applications are used for data analysis. This is partly because production generates a large amount of data from different sources with high potential for AI applications. These applications focus on cost reduction and efficiency improvement. Around 14% of TBIs with AI applications use generative AI (especially LLMs) to understand, summarise and generate text (see also [11]).

Table 1. Number of AI-based systems used in TBIs by category and level [1], [5]

Levels according to [5]	Levels according to [1]	Data analysis [1]	Interaction with customers and employees [1]
Predictive Maintenance	17	22	2
Energy Consumption Forecast	2	2	1
Physical Security	1	2	0
Industrial Automation	27	17	5
Quality Control	5	11	4
Cybersecurity	2	2	0
Process Optimization and Design	7	22	4
Human Machine Interaction	7	11	7

The following section illustrates some specific case studies of how AI methods were implemented and tested for production in the TBIs of AI5production.

Image-based applications

Image-based AI offers an enormously broad range of applications and thus great potential for making production processes and workflows more efficient. This involves both the automation and control of production processes and quality assurance and quality control. A typical example from EDIH AI5production relates to the refurbishment of used photovoltaic modules (PV modules). The increasing number of PV modules also brings with it the problem of recycling. As part of a TBI, the possibility of automated testing of PV modules was investigated. One of the problems was the automated contacting of the module's connectors using a robot. AI-based image processing was used here. The location, position and type of the connector had to be correctly identified and the robot controlled using this data. Without AI methods, this automation step would not be feasible.

AI-supported data analysis

AI-supported data analysis is an important topic for the future, particularly when it comes to detecting anomalies in production (production machines and production facilities). The aim is to ensure quality and reduce waste through early anomaly detection. These approaches are summarised under the term 'predictive maintenance'. Based on data, they allow necessary maintenance to be carried out in a timely and controlled manner in order to prevent unplanned downtime. Particularly in connection with cutting machines (used to remove chips from workpieces), there are repeated fluctuations in quality in fine machining, resulting in scrap. In a TBI, sensor data from the machining centre was collected in real time and processed for the domain experts using AI-based approaches. This made it possible to identify potential improvements for the machining process and reduce scrap (Figure 2).



Figure 2. Aufspannung für eine spanabhebende Bearbeitungsmaschine

Generative AI

The popular generative AI mentioned above offers many possible applications in production, in addition to making everyday office life easier. Two specific examples will be presented.

For manufacturing companies, the purchase of raw materials is an important cost and quality factor. In the case of raw materials that are traded at daily rates on the world market, it is very important for the operating result to purchase the right quantity at the right price at the right time. Global events (weather disasters, political tensions, wars, pandemics, demand forecasts, etc.) influence these world market prices. AI systems offer the possibility of analysing the complex interrelationships in order to derive recommendations for action. In this context, a TBI was carried out to investigate the potential of AI-based steel market analyses for global steel purchasing.

One problem that many manufacturing companies face is that the available machine data cannot be used adequately or at all because there is a lack of interfaces and data scientists. As a fictional example: a machine operator knows that rising temperatures and simultaneously falling pressure in a hydraulic unit are a problem. This would also be evident in the machine data, but no automated alarm trigger can be set without a suitable interface, data analyst and/or programmer. This is where a TBI comes in, in which a tool is to be designed and developed that uses AI to understand an individual problem description in natural language and translate this input into the necessary scripts (R, Python) to automatically generate

evaluations or alarms. This ensures that the knowledge of the employees is optimally stored and utilised. As part of a TBI, an AI prototype was developed that can process natural language input, aggregate and process the necessary data appropriately, and ultimately generate reports or set alarm triggers. This AI assistant facilitates the work of the machine operator and enables their know-how to be captured in the system. At the end of the TBI, a prototype is available that can be used to assess whether further development is beneficial for the company.

Assistance systems

Assistance systems that use AI in combination with other technologies (e.g. augmented reality) to support people in repetitive complex or monotonous work steps play a special role. This can reduce the strain on people and lower the error rate. With the help of augmented reality (AR) and projection systems, markings can be displayed that provide humans with information about the next work step. For example, colour-coded light markings on parts after a manufacturing process based on production plans. The light projection provides humans with information about the part for the next work step (see Figure 3).

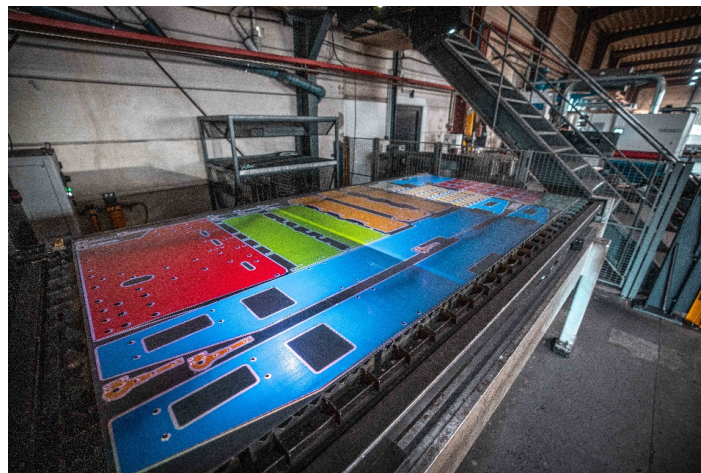


Figure 3. Assistance systems in production – AR application

Another example of assistance systems concerns work preparation. The task here is to clamp components correctly in accordance with specifications for machining. Together with a larger company, the possibilities for using an assistance system were investigated, with the help of which the correct clamping points are projected onto the clamping device. This reduces the work preparation process by 50%. In this project, AI-based systems are integrated and used in multiple ways, for example to recognise individual structural elements and ensure correct projection.

Another area of application is the maintenance and servicing of machines and systems, or even aircraft. In connection with aircraft maintenance, a TBI was carried out in which a concept for the use of an AR-based system was developed. The technology used here is Microsoft HoloLens 2, supported by AI. By integrating a supporting deep learning model, the system is able to independently identify areas and, based on this, make a statement about the actual condition of the system. Effective integration of AR and AI into maintenance processes can have a significant impact on reliability management (i.e., increasing efficiency or accuracy).

Conclusion

Almost half of the TBIs implemented in EDIH AI5production incorporate AI-based methods in various ways. The focus is not on AI technology, but rather on its application in production to automatically solve specific problems. In addition to the general use of generative AI, three areas of application were identified: image-based AI, assistance systems and AI-supported data analysis. This largely corresponds to the current state of research [1]. The TBIs have also shown that this enables companies to automate

complex problems that previously required human intervention. It has also been shown that efficiency and effectiveness can be increased, for example in the form of increased speed and shorter cycle times, lower costs and a reduction in waste and error rates. AI methods are used in an integrated form in many TBIs and are therefore not immediately apparent. Nevertheless, they enable completely new levels of automation and assistance.

For companies in general, but especially manufacturing companies, various recommendations for action can be derived from the results and scientific findings. By categorising the problem, the selection of AI systems for production can be made more accurately and quickly. As already identified in various studies [1], [5], AI-based systems can be successfully introduced and operated in the long term if certain factors are met. These include, on the one hand, an understanding of the technology (e.g. through training) and, on the other hand, piloting (e.g. through TBIs). The focus should be on projects that are based on existing applications and where AI is seen as an extension for problem solving. This leads to greater acceptance of the system, as AI is not perceived as a threat but, in the best case, as a way of making work easier. This is particularly evident in the assistance systems that have brought about improvements for employees. Another important aspect is the consideration of the data situation, as AI-based systems are only as good as the data behind them. Therefore, production data must be collected in advance so that the systems can be trained accordingly. In many cases, retraining is advantageous. The generative AI used in TBIs, especially LLMs, can also be implemented in manufacturing companies in the form of narrow language models, i.e. with smaller, internal company data sets, in order to avoid any biases in freely available or purchased data. These aspects should be taken into account in the decision-making process when AI-based systems are to be used in manufacturing companies.

References

- [1] T. H. Davenport und R. Ronanki, „Artificial intelligence for the real world“, *Harvard business review*, Bd. 96, Nr. 1, S. 108–116, 2018.
- [2] „Generative AI could raise global GDP by 7%“. Zugegriffen: 19. September 2025. [Online]. Verfügbar unter: <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>
- [3] F. Uche Ojika, W. Oseremen Owobu, O. Anthony Abieba, O. Janet Esan, B. Chibunna Ubamadu, und A. Ifesinachi Daraojimba, „The Role of Artificial Intelligence in Business Process Automation: A Model for Reducing Operational Costs and Enhancing Efficiency“, *IJAMRS*, Bd. 4, Nr. 6, S. 1449–1462, Dez. 2024, doi: 10.62225/2583049X.2024.4.6.4046.
- [4] I. M. Enholm, E. Papagiannidis, P. Mikalef, und J. Krogstie, „Artificial Intelligence and Business Value: a Literature Review“, *Inf Syst Front*, Bd. 24, Nr. 5, S. 1709–1734, Okt. 2022, doi: 10.1007/s10796-021-10186-w.
- [5] S. J. Plathottam, A. Rzonca, R. Lakhnori, und C. O. Iloeje, „A review of artificial intelligence applications in manufacturing operations“, *Journal of Advanced Manufacturing and Processing*, Bd. 5, Nr. 3, S. e10159, 2023, doi: 10.1002/amp2.10159.
- [6] S. Feuerriegel, J. Hartmann, C. Janiesch, und P. Zschech, „Generative AI“, *Bus Inf Syst Eng*, Bd. 66, Nr. 1, S. 111–126, Feb. 2024, doi: 10.1007/s12599-023-00834-7.
- [7] N.-A. Perifanis und F. Kitsios, „Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review“, *Information*, Bd. 14, Nr. 2, S. 85, 2023.
- [8] „Article 3: Definitions | EU Artificial Intelligence Act“. Zugegriffen: 19. September 2025. [Online]. Verfügbar unter: <https://artificialintelligenceact.eu/article/3/>
- [9] „Künstliche Intelligenz (KI) und maschinelles Lernen - Fraunhofer IKS“, Fraunhofer-Institut für Kognitive Systeme IKS. Zugegriffen: 19. September 2025. [Online]. Verfügbar unter: <https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html>
- [10] J. McCarthy, M. L. Minsky, N. Rochester, und C. E. Shannon, „A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955“, *AI Magazine*, Bd. 27, Nr. 4, S. 12–12, Dez. 2006, doi: 10.1609/aimag.v27i4.1904.
- [11] F. Sümeecz, „AI5production - Der Weg zur digitalen Transformation in der Produktion“. TU Wien, 2025.

Digital data consistency for drive technology in building services engineering

Gerhard Zucker
AIT Austrian Institute of Technology
gerhard.zucker@ait.ac.at

Tanja Ivanovic
AIT Austrian Institute of Technology
tanja.ivanovic@ait.ac.at

Kurzzusammenfassung

Das Projekt DRIVE_THRU zielte auf die Schaffung einer durchgängigen digitalen Datenbasis für Komponenten der technischen Gebäudeausrüstung (TGA), insbesondere Antriebstechnik, ab. Auf Basis offener Standards wie IFC (Industry Foundation Classes), dem offenen Format für Building Information Modeling (BIM), wurden Methoden entwickelt, um Herstellerdaten nahtlos in Planungs-, Ausführungs- und Betriebsprozesse zu integrieren. Die Ergebnisse umfassen eine BIM-konforme Revit-Demofamilie mit Parametern für Lebenszyklusanalyse (LCA) und Kreislaufwirtschaft sowie einen prototypischen Workflow zur Datenweitergabe über den gesamten Lebenszyklus. Die dadurch gewonnene Datenkonsistenz und Transparenz unterstützt fundierte Entscheidungen in Planung, Ausführung und Betrieb und leistet einen Beitrag zu nachhaltigem Bauen.

Keywords: Building Information Modeling (BIM), Gebäudetechnik, Antriebstechnik, Kreislaufwirtschaft, Life Cycle Assessment (LCA)

Abstract

The DRIVE_THRU project aimed to create a comprehensive digital database for technical building equipment (TBE) components, particularly drive technology. Based on open standards such as IFC, methods were developed to seamlessly integrate manufacturer data into planning, execution and operational processes. The results include a BIM-compliant Revit demo family with parameters for life cycle analysis (LCA) and circular economy, as well as a prototypical workflow for data transfer throughout the entire life cycle. The resulting data consistency and transparency supports informed decisions in planning, execution and operation and contributes to sustainable construction.

Keywords: Building Information Modeling (BIM), Life Cycle Assessment (LCA)

Introduction

The construction industry is striving to make great strides in terms of digitalization and sustainability. Technical building equipment (TBE) plays a central role in the energy efficiency and operational quality of buildings, but it is precisely in this area that there is often a lack of consistent, structured data availability throughout the entire life cycle. Although manufacturers are increasingly providing digital product information, this is often incompatible with planning tools or unusable for later phases of use. Planning is often based on incomplete or redundant information, leading to inefficiencies and media breaks. The DRIVE_THRU project addresses this problem using the example of drive technology components and is developing a prototypical data model that makes information available throughout the entire process, from production to planning and tendering to operation, with a particular focus on the integration of sustainability aspects such as life cycle data and circular economy.

Initial situation and objectives

The project partners identified a key challenge in the lack of interoperability and semantic consistency of technical product data in BIM processes. Drive components must be modeled as generic components in the early stages of building design, with essential technical, environmental, and operational information either missing or only added later. This discontinuity leads to information loss and increased coordination effort. The aim of the project was to develop a uniform, open data structure that maps both

generic and product-specific information and makes it usable in common software environments. The use of open standards such as IFC [1] was intended to lay the foundation for broader implementation in the sector. In addition, the aim was to integrate selected environmental indicators in the form of LCA parameters [2] into the model in order to systematically incorporate ecological aspects into the planning process.

Methodology

The methodological implementation focused on the systematic analysis of information requirements throughout the life cycle of technical components. To this end, interviews were conducted with experts from the fields of planning, execution, and building operation in order to identify the specific information requirements in each phase. At the same time, environmental product declarations (EPDs) were analyzed to extract reliable environmental indicators. These findings were used to derive a consistent set of parameters that includes both operational and ecological information. These parameters were integrated into a prototypical Revit family that serves as a demonstrator and shows what structured data modeling for drive technology can look like. The technical implementation took into account the IFC schema in accordance with ISO 16739 to ensure maximum interoperability. In addition to modeling the parameters, an exemplary workflow was developed that illustrates the transfer and use of this data in various software environments and application phases.

Results and demonstrator

A key result of the project is the development of a standardized, BIM-compatible Revit family for selected components from our project partner Belimo Automation. This family contains a wide range of parameters that reflect both technical characteristics such as energy consumption, connection values, and media compatibility, as well as ecological indicators such as global warming potential (GWP), recyclability, and service life. In addition, operationally relevant data such as serial numbers and maintenance cycles were taken into account to increase usability in facility management. The demonstrator shows not only the structured parameterization, but also the transfer points between the individual life cycle phases, such as the transfer of generic tender data to manufacturer-specific planning models. The prototypical workflow illustrates how this data can be automatically transferred, further processed, and ultimately used for operational purposes. This establishes a digital chain that extends from the initial planning idea to building operation. Figure 1 shows the Belimo Energy Valve, which was used as a reference for the demonstrator.

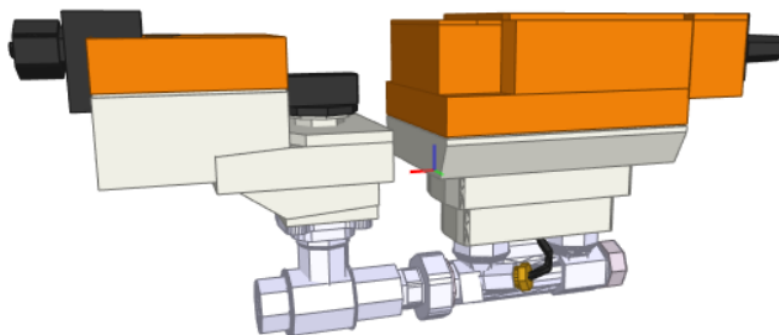


Figure 1. Belimo Energy Valve

Supplementary analyses: life cycle, scenarios, and material effects

A key element of the DRIVE_THRU project was an in-depth examination of the ecological impact of technical components throughout their entire life cycle. The focus was on the life cycle assessment (LCA) of an exemplary product – specifically the Energy Valve EVO15R2+BAC. The aim was to evaluate different life cycle scenarios in terms of their carbon footprint and resource use. To this end,

two variants were compared: the “cradle to grave” scenario, which maps the classic linear life cycle from raw material extraction to disposal, and the “cradle to cradle” scenario, which focuses on material cycles and reuse.

The calculations were based on recognized standards such as EN 15804 [3] and ISO 14025 [4] and took into account both embodied carbon and emissions during the use phase (operational carbon). Assuming a building life of 100 years and a product life of 10 years, the component under consideration had to be replaced a total of nine times. In the cradle-to-grave model, the material requirements were entirely made up of primary material, while in the cradle-to-cradle approach, a large part could be covered by secondary material. This led to a significant reduction in the ecological footprint. It was found that 78% of the product mass could be recycled and the remaining 22% could be used for energy recovery – without any material having to be sent to landfill.

These quantitative results were used to develop a concrete understanding of the relevance of ecological parameters in digital modeling. The underlying material data and sustainability indicators were transferred directly into the IFC data structure and embedded as extended parameters in the prototypical Revit family. This not only provided evidence of the product's environmental performance, but also created the technical prerequisite for automatically integrating LCA indicators into digital planning and decision-making processes.

Innovation and benefits for industry

The DRIVE_THRU project is making a significant contribution to the digitalization of the building services engineering industry, particularly with regard to data availability and sustainability integration. The combination of open standards, manufacturer data, and environmental parameters is a novelty, as previous data models were either technically or ecologically focused, but rarely able to integrate both. For manufacturers, the structured preparation of their product data represents a competitive advantage, as it allows them to be integrated into digital planning processes and given preferential consideration by planners. Planners benefit from improved data availability and can make informed decisions regarding sustainability and operation at an early stage. Finally, operators receive complete, machine-readable product information, which forms the basis for data-based facility management. The findings gained in the project will flow directly into a planned follow-up project entitled “TGA Connect,” which will promote the standardization of these parameters at the national level.

Summary

DRIVE_THRU demonstrates how consistent digitization and the use of open standards can create a continuous flow of information for building services components. The structured modeling of technical and ecological parameters in a Revit family and the development of an interoperable workflow show in a practical way how manufacturer data can be made usable for planning and operation. The combination of data modeling, sustainability considerations, and standardization makes a relevant contribution to the digital transformation and ecological optimization of the building sector.

Acknowledgments

We would like to thank our project partner Belimo Automation Handelsgesellschaft m.b.H. for their excellent cooperation in the project and their support in developing the project results.

References

- [1] ISO International Organization for Standardization, *EN ISO 16739:2021 - Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries*. 2021. Zugriffen: 10. April 2020. [Online]. Verfügbar unter: <https://www.austrian-standards.at/de/shop/din-en-iso-16739-1-2021-11~p2602356>
- [2] *BS EN 15978:2011 Sustainability of construction works. Assessment of environmental performance of buildings. Calculation method*. Zugriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://www.en-standard.eu/bs-en-15978-2011-sustainability-of-construction-works-assessment-of-environmental-performance-of-buildings-calculation-method>
- [3] *ÖNORM EN 15804: Nachhaltigkeit von Bauwerken - Umweltproduktdeklarationen - Grundregeln für die Produktkategorie Bauprodukt*. Zugriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://www.austrian-standards.at/de/shop/onorm-en-15804-2022-02-15~p2614831>
- [4] A. Standards, *ISO 14025: Umweltkennzeichnungen und -deklarationen - Typ III Umweltdeklarationen - Grundsätze und Verfahren*. Zugriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://www.austrian-standards.at/de/shop/onorm-en-iso-14025-2010-07-01~p1810683>

Efficiency improvements in WIG and plasma welding inverters

Conrad Gstöttner
Austrian Institute of Technology GmbH
Conrad.Gstöttner@ait.ac.at

Andreas Pfeffer
Austrian Institute of Technology GmbH
Andreas.Pfeffer@ait.ac.at

Peter Jonke
Austrian Institute of Technology GmbH
Peter.Jonke@ait.ac.at

Katharina Machtinger
Austrian Institute of Technology GmbH
Katharina.Machtinger@ait.ac.at

Kurzzusammenfassung

Aufgrund des hohen Leistungsbedarfs von Schweißinvertern ist die Energieeffizienz ein wichtiges Qualitätsmerkmal dieser Geräte, weswegen Energieeffizienzmaßnahmen im Fokus dieses TBI standen. Im Rahmen des Projektes wurde moderne Laborinfrastruktur des Austrian Institute of Technology (AIT) verwendet, um Messungen an einem Schweißinverter durchzuführen, der dem aktuellen Industriestandard entspricht. Auf Basis dieser Messungen, sowie Literatur- und Patentrecherchen wurden Verbesserungspotentiale aufgezeigt und Möglichkeiten diskutiert diese auszuschöpfen. Aufgrund des Reifegrades der aktuellen Technologie ist der größte Hebel durch einen Umstieg auf moderne Wide-Bandgap-Halbleiter in Kombination mit einem entsprechend angepasstem Schaltungsdesign und innovativen Regelungskonzepten gegeben. Allerdings ist die bisher geringe Verbreitung dieser Technologien in diesem Anwendungsgebiet ein großes Hemmnis für die Industrie. Daher wurden im Rahmen dieses TBI moderne digitale Möglichkeiten wie Hardware-in-the-Loop-Tests (HIL-Tests) aufgezeigt, welche den Entwicklungsprozess beschleunigen können.

Keywords: Schweißinverter, Leistungselektronik, Wide-Bandgap-Halbleiter, Effizienzsteigerung

Abstract

The high electrical energy demand of welding inverters makes their energy efficiency a key quality feature, which is why measures to increase the energy efficiency were of primary interest for this TBI. Within the project, modern laboratory infrastructure of the Austrian Institute of Technology (AIT) was used to take measurements on an industrial-standard welding inverter. Based on these measurements, as well as literature and patent reviews, potentials for raising the energy efficiency were highlighted and possibilities for exploiting them were discussed. Due to the maturity of the current welding inverter technology, the greatest potential lies in transitioning to modern wide-bandgap semiconductors with accordingly adapted circuit designs and innovative control concepts. However, the limited adoption of these technologies in the field of welding inverters poses a barrier for the industry due to prolonged development cycles. Therefore, within the scope of this TBI modern digital methods such as Hardware-in-the-Loop (HIL) testing were demonstrated to accelerate the development process.

Keywords: welding inverters, power electronics, wide-bandgap-semiconductors, increase in energy efficiency

Introduction

According to EU Directive 2019/1784, welding inverters with a projected final energy consumption of 6 TWh for 2030, corresponding to 2.4 million tons of CO₂ equivalents, are classified as energy-related products [1]. Among other things, the directive stipulates minimum requirements for the energy efficiency of welding inverters, for example 85% efficiency and a maximum power consumption of 50W in idle mode for devices with three-phase power supply and DC output. With the help of modern wide-bandgap semiconductors such as silicon carbide (SiC) semiconductors, which are already state of the art in electromobility and achieve efficiencies of approx. 98-99% (see, for example, [2], [3]), these minimum requirements are likely to be significantly exceeded. Another advantage of improved

efficiency is the reduction of waste heat, which optimizes the cooling of the device and, in the ideal case, allows it to be passive. This further increases overall efficiency and reduces idle power consumption. Lower waste heat also enables more compact designs and, consequently, smaller housings, which reduces the resources required for these as well as for transport. Due in particular to the longevity of welding inverters, which is typically 10 years in harsh industrial environments, it is important to develop devices with maximum energy efficiency and compact design as early as possible.

In industrial practice, devices are currently in widespread use that are predominantly manufactured in the Far East using standard components. Although these devices meet the minimum requirements currently stipulated by the EU, they generate little added value in Europe and fail to exploit their full potential, particularly in terms of energy efficiency, but also with regard to network interference. In terms of efficiency, switching to modern semiconductor technologies, such as the use of SiC semiconductors, is the only way to achieve substantial improvements. However, this requires more than just replacing the semiconductor switching elements; a new design is necessary to fully exploit the potential [4]. This is a particular obstacle for the industry, as these technologies have hardly been used in this area of application to date. With regard to grid interactions, modern control methods and high-precision control of the individual switching elements are particularly promising options for specifically reducing the presence of relevant harmonics.

Problem-solving approach / Procedure

As part of the TBI, modern laboratory infrastructure at AIT was used to carry out measurements on a welding inverter that complies with the current industry standard. The welding inverter used for the measurements is suitable for both tungsten inert gas (TIG) and plasma welding. The efficiency and network effects were measured at various operating points for TIG and plasma welding. The measurements, in combination with literature and patent searches, served to identify potential improvements and ways to exploit them. Figure 2 below shows examples of the efficiency and the active and apparent power measured on the input side for the plasma operating points measured.

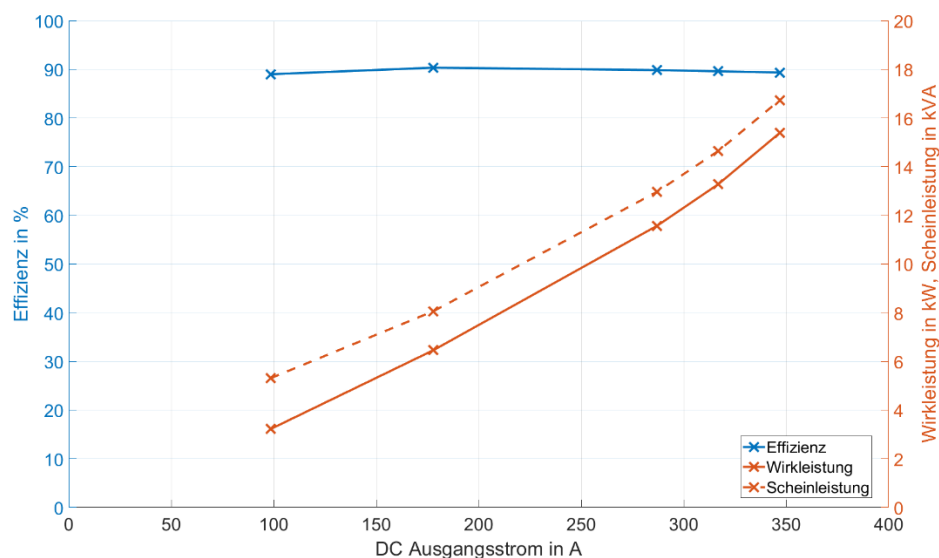


Figure1. Measured efficiency, active power, and apparent power at various operating points for plasma welding.

At all operating points shown for plasma welding, the efficiency is approximately 90%. The ratio of active to apparent power improves from approximately 0.61 at the operating point near 100A output current to approximately 0.92 at the operating point near 350A output current. Measurement of the network effects also showed that the input current curves deviate from a purely sinusoidal curve at all operating points. In particular, 5th and 7th harmonics occur, as well as isolated higher odd harmonics

with significant amplitudes. The manufacturer of the welding inverter was provided with the complete measurement results..

The measurement results in Figure 2 show that efficiencies in the range of 90% can be achieved with the technologies currently used in welding inverters, such as Si-MOSFET and Si-IGBT switching elements. However, compared to applications such as electromobility, where modern wide-bandgap semiconductor elements are used to achieve efficiencies of 98-99%, there is clear potential for improvement. The advantages of SiC MOSFETs over Si IGBTs in inverters in the field of electromobility have been investigated, for example, in [2], [5], [6], and [7]. In the field of welding inverters, efficiency increases of a similar magnitude are likely to be achievable. Current research on the use of SiC semiconductors in welding inverters can be found, for example, in [8] and the references cited therein.

With regard to power factor, the measurement results indicate significant potential for improvement, particularly in the partial load range. A reduction in the harmonics of the current curves measured on the input side would be desirable at all operating points in order to be prepared for future legal guidelines on network effects. To improve network interference across the entire load range, the passive input filter used in the measured inverter could be replaced by an active power factor correction (PFC) module. A detailed analysis of various circuits for active power factor correction can be found, for example, in [9], [10].

Results

The measurement and research results presented in excerpts in the previous section, as well as the resulting potential for improvement and opportunities to exploit this potential, were prepared, presented, and discussed with the project partner. The development of a new generation of welding inverters that are superior to the current state of the art in terms of energy efficiency, grid impact, and compactness is a desirable goal for the project partner and the welding industry as a whole. Due to the maturity of the technology currently in use, a switch to modern wide-bandgap semiconductors with appropriately adapted circuit design and innovative control concepts is essential. For the project partner, representing many manufacturers of welding inverters, this represents an obstacle, as these modern technologies have so far mainly been used in prototypes in this area of application. In the literature, this prototypical use is only found in inverters with DC output. However, for flexible welding inverters that are intended to cover a wider range of applications, various AC signal forms are also required at the output. This raises the research question of how the topology for this application could be realized with the new switching element technologies. Im Gegensatz zu der Verwendung von Wide-Bandgap-Halbleitern ist der Einsatz einer aktiven PFC bei Schweißinvertern bereits häufig anzutreffen. Dies dient der Verbesserung des Leistungsfaktors und der Reduktion von Oberschwingungen in den eingangsseitigen Stromverläufen. Der Einsatz einer aktiven PFC eröffnet außerdem neue Möglichkeiten hinsichtlich der Regelung des Gesamtsystems, welches aus PFC, Hauptinverter und der für Inverter mit AC-Ausgang benötigten Schaltung zur Ausgangssignalformung besteht. Anstatt die einzelnen Baugruppen regelungstechnisch getrennt voneinander zu betrachten, könnte eine gesamtheitliche übergeordnete Regelung der Teilsysteme weiteres Potential für Effizienzsteigerungen bieten.

The use of modern semiconductor elements, such as SiC MOSFETs, results in higher component costs compared to currently used technologies. Although these costs are offset by higher efficiency over the operating life, they initially result in an increase in the price of the device. However, by taking a holistic view of all the assemblies typically used in a welding inverter, synergy effects can also be exploited to reduce the number of switching elements required. One open research question in this context is the reduction of switching elements by integrating the pilot inverter, which was previously constructed separately, into the main circuit board. In some applications, this is necessary to establish and maintain the arc.

The result of the patent research carried out as part of the TBI is that the development of a new generation of welding inverters with the above-mentioned innovations

- use of wide bandgap semiconductors,
- use of active PFC,
- (partial) integration of the pilot inverter or its function into the main inverter, and
- implementation of a higher-level control system for the entire system instead of separate control of the individual assemblies,

is not subject to any significant restrictions due to existing patents.

However, the research and development costs associated with such a new development are difficult to manage on their own, especially for SMEs. In addition to technical expertise in power electronics design with wide-bandgap semiconductors, the creation of simulation models, and digital control, modern laboratory infrastructure is also required. Various options for accelerating the development process were discussed, including hardware-in-the-loop (HIL) testing. These are seen as a way of testing control algorithms and topologies or circuit designs at an early stage of development in a simulated but realistic environment. The hardware on which the control algorithms are executed is coupled with a real-time system that emulates the systems to be controlled (PFC, main inverter, etc.), see Figure 3.

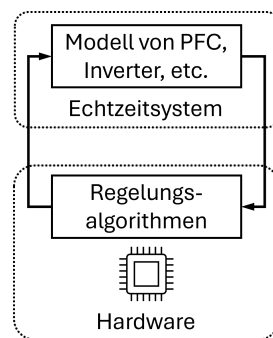


Figure 2. Schematic representation of a HIL test environment.

This method enables hardware and algorithms to be validated early and reliably without the need for the physical target system and without the risk of damage to existing real-world structures in the event of software errors [11]. It also reduces the time and resources required to build different prototypes. The project can also benefit from modern development practices such as clean code, design patterns, modularization, and the implementation of a continuous integration and continuous deployment (CI/CD) pipeline with automated unit tests [12], [13].

Summary

Measurements carried out as part of the TBI on a welding inverter, as often used in industry, have shown that there is potential for improvement in terms of energy efficiency and network effects. In particular, the wide range of applications for welding inverters in the EU mentioned at the outset highlights the importance of reducing emissions, which goes hand in hand with efficiency improvements in welding inverters. In order for manufacturers to be prepared for future legal guidelines on energy efficiency and network effects, a switch to modern wide-bandgap semiconductors, which are already state of the art in other industrial fields, and the use of active PFC is recommended. In addition, the development of an innovative control concept for the entire system, consisting of PFC, main inverter, pilot inverter function, and output signal shaping circuitry, offers potential for efficiency gains. The use of modern digital methods, such as hardware-in-the-loop tests, can accelerate the development process. Within the framework of the TBI, the basis for further research projects in cooperation with industry partners was created in order to put the identified improvement potential and the discussed possibilities for exploiting this potential into practice. From today's perspective, the implementation of the proposed solutions could pave the way for technological leadership in the field of welding inverters and significantly reduce the overall energy consumption in this sector.

Acknowledgements

The authors would like to thank SBI GmbH, in particular J. Niedermayer and F. Spulak, for providing an industry-standard welding inverter and peripherals for carrying out the measurements, as well as for the technical exchange.

References

- [1] Verordnung (EU) 2019/1784 der Kommission vom 1. Oktober 2019 zur Festlegung von Anforderungen an die umweltgerechte Gestaltung von Schweißgeräten gemäß der Richtlinie 2009/125/EG des Europäischen Parlaments und des Rates. Zugegriffen: 25. Juli 2025. [Online]. Verfügbar unter: <https://eur-lex.europa.eu/eli/reg/2019/1784/oj>
- [2] B. Shi, A. I. Ramones, Y. Liu, H. Wang, Y. Li, S. Pischinger, und J. Andert, „A review of silicon carbide MOSFETs in electrified vehicles: Application, challenges, and future development“, IET Power Electronics, Bd. 16, Nr. 12, S. 2103–2120, Sep. 2023, doi: 10.1049/pel2.12524.
- [3] R. Unseld, „I estimate that 70 % of the SiC demand will account for the e-mobility sector“, ATZ Electronics Worldwide, Bd. 19, Nr. 6, S. 14–17, Juni 2024, doi: 10.1007/s38314-024-1886-9.
- [4] S. Acharya, N. S. Chauhan, und S. K. Mishra, „Replacing Si-IGBT by SiC Mosfet in high Gain Inverter: Challenges and Opportunities“, in Tagungsband der 2018 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Chennai, India: IEEE, Dez. 2018, S. 1–6. doi: 10.1109/pedes.2018.8707565.
- [5] S. Hain, M. Meiler, und M. Denk, „Evaluation of 800V Traction Inverter with SiC-MOSFET versus Si-IGBT Power Semiconductor Technology“, in Tagungsband der PCIM Europe 2019; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management, Nürnberg, Deutschland, Mai 2019, S. 1–6.
- [6] M. Nitzsche, C. Cheshire, M. Fischer, J. Ruthardt, und J. Roth-Stielow, „Comprehensive Comparison of a SiC MOSFET and Si IGBT Based Inverter“, in Tagungsband der PCIM Europe 2019; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management, Nürnberg, Deutschland, Mai 2019, S. 1–7.
- [7] X. Ding, M. Du, T. Zhou, H. Guo, C. Zhang, und F. Chen, „Comprehensive Comparison between Sic-mosfets and Si-igbts Based Electric Vehicle Traction Systems under Low Speed and Light Load“, Energy Procedia, Bd. 88, S. 991–997, Juni 2016, doi: 10.1016/j.egypro.2016.06.124.
- [8] A. Marneni, N. Babu, S. K. Subramaniam, P. Murugesan, und J. J. Prakash, „Design, analysis and testing of three-phase isolated SiC-based high current DC-DC converter for arc welding power supply“, Electr Eng, Apr. 2025, doi: 10.1007/s00202-025-03089-5.
- [9] J. W. Kolar und T. Friedli, „The Essence of Three-Phase PFC Rectifier Systems—Part I“, IEEE Transactions on Power Electronics, Bd. 28, Nr. 1, S. 176–198, Jan. 2013, doi: 10.1109/tpel.2012.2197867.
- [10] T. Friedli, M. Hartmann, und J. W. Kolar, „The Essence of Three-Phase PFC Rectifier Systems—Part II“, IEEE Transactions on Power Electronics, Bd. 29, Nr. 2, S. 543–560, Feb. 2014, doi: 10.1109/tpel.2013.2258472.
- [11] H. P. Nguyen, T. T. Nguyen, M. P. Le, und T. N. Tran, „Experimental platforms consisting of software-in-the-loop, hardware-in-the-loop and power-hardware-in-the-loop for developing MPPT charge controller for photovoltaic system“, Solar Energy, Bd. 298, S. 113666, Sep. 2025, doi: 10.1016/j.solener.2025.113666.
- [12] A. Barbie, W. Hasselbring, und N. Pech, „Continuous Integration Testing of Embedded Software with Digital Twin Prototypes“, 26. Juni 2021, TechRxiv. doi: 10.36227/techrxiv.14770983.v1.
- [13] P. Cassieri, M. Esposito, S. Romano, und G. Scanniello, „On the use of Test-Driven Development for Embedded Systems“, Information and Software Technology, Bd. 187, S. 107779, Nov. 2025, doi: 10.1016/j.infsof.2025.107779.

Multi-sensor imaging for the qualitative characterization of used solar modules

Alexander Eckert
Austrian Institute of Technology
Alexander.eckert@ait.ac.at

Georg Kaniak
Austrian Institute of Technology
Georg.kaniak@ait.ac.at

Philipp Jelke
Austrian Institute of Technology
Philipp.jelke@ait.ac.at

Laurin Ginner
Austrian Institute of Technology
Laurin.ginner@ait.ac.at

Kurzzusammenfassung

Die kosteneffiziente Qualitätsprüfung gebrauchter Solarmodule für eine mögliche Wiederverwendung wird angesichts stark steigender Rücklaufmengen zunehmend wichtiger. Ein niederösterreichisches Startup beauftragte das Austrian Institute of Technology (AIT) daher mit der Entwicklung eines vollautomatisierbaren Multi-Sensor-Imaging-Inspektionssystems. Ziel war die Konzeption und Erprobung einer optischen Inspektionsstation. Dafür wurden verschiedene Beleuchtungssysteme, Kameras und Objektive unter realen Bedingungen getestet und verglichen. Auf dieser Basis wurden aufgabenoptimierte Systeme implementiert, die nun die Grundlage für eine automatisierte Qualitätsbewertung gebrauchter Solarmodule bilden.

Keywords: Kreislaufwirtschaft, Optische Inspektion

Abstract

The cost-efficient quality assessment of used PV modules for potential reuse is becoming increasingly important due to the expected surge in return volumes. To address this, a Lower Austrian startup commissioned AIT to develop a fully automatable multi-sensor imaging inspection system. The project aimed to support the design and validation of an optical inspection station. For this purpose, various lighting systems, cameras, and lenses were tested and compared under real-world conditions. Based on these evaluations, task-optimized systems were implemented, forming the foundation for automated quality assessment of used PV modules.

Keywords: Circular economy, optical inspection

Introduction

More than 4 billion solar modules (photovoltaic modules) have now been installed worldwide—a billion were added last year alone. Accordingly, a sharp increase in the return of decommissioned photovoltaic modules (PV modules) is expected in the coming years. Many of these are not defective, but are being replaced as part of repowering measures, even though they often still achieve efficiencies of 80–90% and only show annual performance losses of 1–2% [1]. In Germany alone, around 50 million used modules are expected to require processing by 2030 [2]. This development highlights the urgent need for efficient reuse and recycling strategies – and at the same time opens up opportunities for automated solutions in the field of circular economy.

Current quality checks on used PV modules are mainly carried out manually. Since visual and electrical inspections are performed by humans, they are time-consuming, inconsistent, and difficult to scale – and thus represent a major bottleneck in the reuse process. Despite the high market demand, the degree of automation in module inspection has been low to date. At present, almost all returns are visually inspected, but a large proportion are not subjected to any further inspections [3]. The majority are currently sent for material recycling without precise individual characterization [4].

To close this gap, the Lower Austrian start-up initiated a development project together with AIT. The aim was to replace manual inspections with camera-based, automated inspection systems. The focus was on developing a robust, high-throughput optical inspection solution that reliably detects defects, reduces manual effort, and thus enables scalable reuse options.

Problem-solving approach / Procedure

To develop a robust, application-specific inspection system, AIT first analyzed typical used PV modules, common defect patterns, and how they could be detected [5]. Relevant damage patterns included cracks, inactive cells, delamination, and blistering on the back.

Taking into account mechanical constraints, high processing speeds, and the necessary optical resolution, a customized inspection system was designed. This combines several camera and lighting configurations:

- Front side inspection: High-resolution electroluminescence (EL) cameras [6] for detecting electrical and structural defects, supplemented by RGB cameras for visual surface evaluation.
- Back side inspection: RGB cameras in combination with line lights and directional lighting to highlight surface textures and detect delamination or blistering.

The resulting multi-sensor configuration enables task-optimized, automated inspection with high throughput and was specifically developed for the requirements of the planned inspection line.

Results

The project successfully developed and tested a multi-camera inspection system and tailored it to the requirements of an automated PV module inspection line. The evaluation showed that different inspection criteria require the use of different cameras, lenses, and lighting systems. The selected components were validated through testing and then recommended for integration, see Figure 1.

The system is now in use in the client's production environment and enables reliable high-speed inspection. This represents an important step toward fully automated quality testing of used PV modules.

This result not only strengthens the technological basis of the circular economy in the photovoltaic sector, but can also be transferred to other players in Europe and beyond. In addition, the multi-sensor approach offers potential for expansion to other complex material flows in the field of recycling.

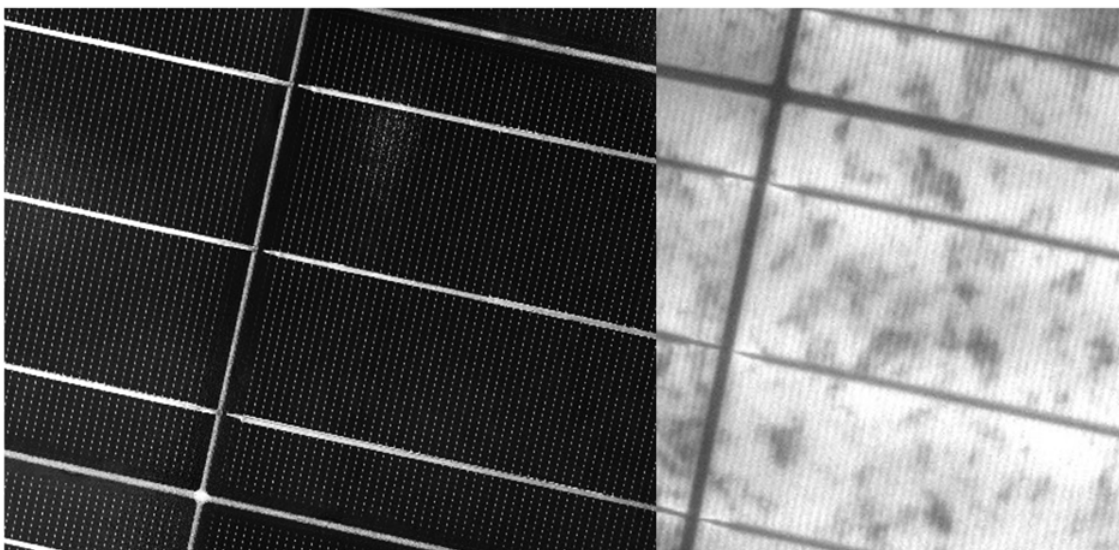


Figure 1: Polycrystalline solar cell imaged with different camera systems (left: RGB, right: monochrome, electroluminescence)

Summary

As billions of PV modules approach the end of their life cycle, efficient reuse is becoming increasingly important. Many decommissioned modules are still functional, but current manual testing methods are too slow and labor-intensive. Together with AIT, the Lower Austrian start-up has therefore developed an automated multi-camera inspection system. By combining EL and RGB cameras with specialized lighting, it enables fast and reliable quality assessment. The system is already integrated into the inspection line and makes a decisive contribution to the scalability of reuse and the implementation of circular economy models in the photovoltaic sector.

References

- [1] Santos, Sofia & Torres, João Paulo & Fernandes, Carlos & Lameirinhas, Ricardo. (2021). The impact of aging of solar cells on the performance of photovoltaic panels. *Energy Conversion and Management*: X. 10. 100082. 10.1016/j.ecmx.2021.100082.
- [2] IRENA and IEA-PVPS (2016), “End-of-Life Management: Solar Photovoltaic Panels,” International Renewable Energy Agency and International Energy Agency Photovoltaic Power Systems, ISBN 978-92-95111-99-8
- [3] Umweltbundesamt (2024, 26. März). Photovoltaik – Behandlung und Recycling. Abgerufen von Umweltbundesamt-Webseite
- [4] IEA PVPS Task 12 (2022). Status of PV Module Recycling in Selected IEA PVPS Task 12 Countries. International Energy Agency Photovoltaic Power Systems Programme
- [5] Maziuk, Maurycy & Jasińska, Laura & Domaradzki, Jaroslaw & Chodasewicz, Paweł. (2023). Imaging methods of detecting defects in photovoltaic solar cells and modules: a survey. *Metrology and Measurement Systems*. 30. 10.24425/mms.2023.146426.
- [6] Frazão, M. & Silva, José & Lobato, Killian & Serra, J.M.. (2016). Electroluminescence of silicon solar cells using a consumer grade digital camera. *Measurement*. 99. 10.1016/j.measurement.2016.12.017.

Digitization of an energy efficiency concept with reporting function

Clemens Schwaiger
CDP Center for Digital Production GmbH
clemens.schwaiger@acd-p.at

Sascha Gent
CDP Center for Digital Production GmbH
sascha.gent@acd-p.at

Kurzzusammenfassung

Das Projekt konzentrierte sich auf die Überwachung des Energieverbrauchs in einem Walzwerk. Es ermöglicht eine Analyse des Energieaufwands pro Produkt und je Gewicht an eingesetztem Rohmaterial. Zudem werden die Energieverbräuche in drei Betriebszuständen der Produktionsanlagen eingeteilt und auf einem Dashboard visualisiert. Dieses dient sowohl der Verbesserung der Energieeffizienz als auch der Fehlererkennung, indem es Anomalien im Verbrauch aufzeigt und frühzeitig auf technische Störungen hinweist. So können Produktionsprozesse optimiert und Ausfallzeiten minimiert werden, wodurch eine nachhaltige und kosteneffiziente Fertigung unterstützt wird. Die Ergebnisse werden den Anwendern aufbereitet zur Verfügung gestellt und in ein bestehendes Supervisory Control and Data Acquisition (SCADA) - System integriert.

Keyword: Digitalisierung, Energieeffizienz, Kalanders

Abstract

The project focused on energy monitoring in a rolling mill by enabling the analysis of energy consumption per product and per weight of raw material used. Furthermore, this energy consumption is divided into three operating states of the production facilities and visualized on a dashboard. This serves both to improve energy efficiency and detect errors by highlighting consumption anomalies and providing early warning of technical malfunctions. It allows production processes to be optimized and downtime minimized, thus supporting sustainable and cost-efficient production. The results are processed and made available to users and integrated into an existing SCADA system.

Keyword: Digitalization, energy efficiency, Kalenders

Introduction

This article deals with a rolling mill consisting of several systems for smoothing and compacting rubber. These “calenders” are a central component in the production of sealing material. They basically consist of two polished and heated rollers that form a variable gap. The material is passed through this gap in several passes to produce a mat with the desired thickness.

These calenders are supplied with electrical energy to operate the drives and with steam to regulate the temperature of the rollers. The peripherals include an extraction system and a solvent recovery system, which require additional electrical energy and steam. An overview of the energy flows in and around the rolling mill is shown in Figure 1.

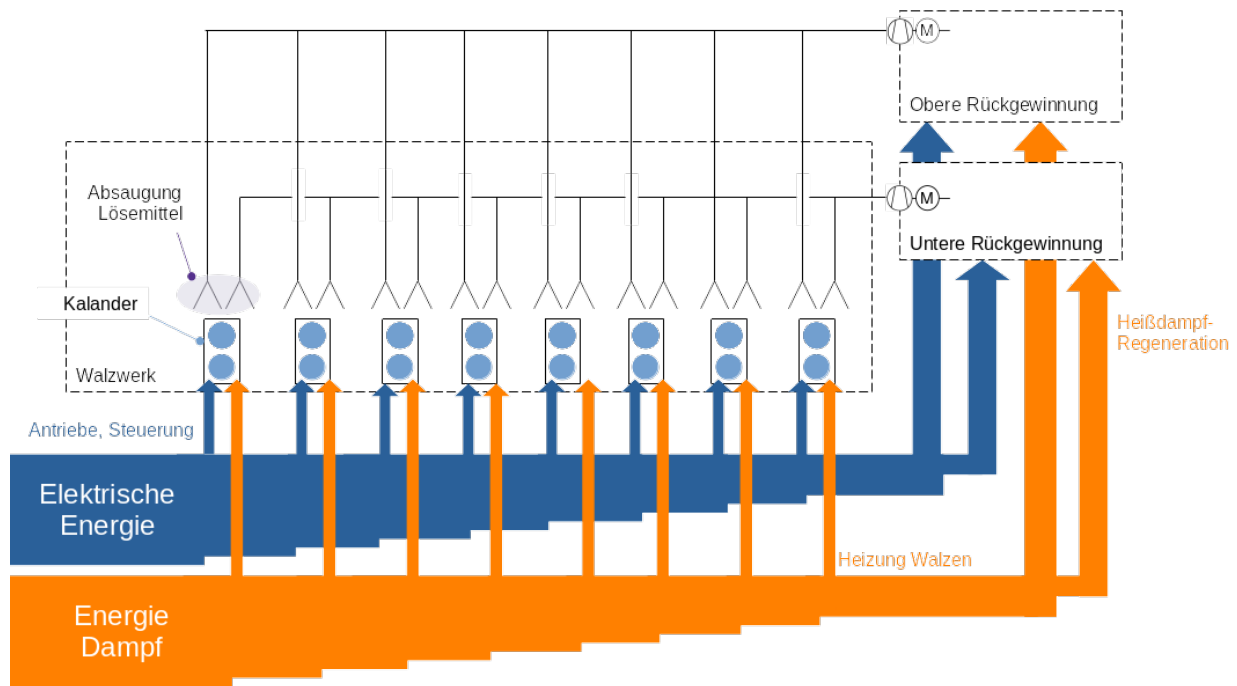


Figure 1. Energy flows in the rolling mill area

These calenders are energy-intensive production machines, particularly due to their consumption of steam, which is produced by burning fossil fuels. Optimizing energy consumption is therefore a significant cost factor and also has a positive effect on the CO₂ balance.

Energy monitoring and management, e.g., in the form of legally binding management systems such as ISO 50001, form the basis for improvement measures [1]. However, such structures must be expanded to enable truly strategic energy management [2]. Specifically, this means harmonizing and contextualizing energy data with other production data to create an informative database. Only in this way can the necessary information content be provided for decision-makers and constructive analyses. Prior to this project, such a database did not yet exist in the company, which is why energy consumption was largely opaque apart from absolute values, making optimizations difficult to implement. Products available on the market (e.g., FourJaw, iFactory, and similar) cannot adequately map the very specific production process or, like the market leader EnMS from ABB, are too complex to be integrated into an SME. This would result in personnel overhead that would not be offset by the targeted savings. For this reason, a customized solution that could also communicate with the company's existing systems was the only solution. This solution for determining and processing the data required for meaningful and robust energy monitoring is the subject of the project on which this article is based.

Problem-solving approach / Procedure

Contextualization is specific to the respective production and must be based on the types of energy forms, equipment, materials, and operating strategies. Different approaches can be pursued here. In the case of this rolling mill, the following key figures should be derived:

- Energy consumption in relation to the weight of the material used
- Energy consumption in relation to individual products
- Energy consumption per operating state of the plant
- Breakdown of energy forms

Energy consumption per operating state requires the definition of the operating state. An abstract and universal approach is to classify the states according to their added value:

- a. Directly value-adding: Energy input directly required for the production process (e.g., energy to drive the rolls during the rolling process)
- b. Indirectly value-adding: Other energy consumption required for the production processes. (e.g., energy to operate the fans for extracting solvent vapors)
- c. No value creation: (Theoretically) avoidable states (e.g., operation of the hydraulic unit to maintain system pressure during (longer) periods without production cycles)

The required energy and production data were consolidated in the in-house SCADA system and the necessary calculations were performed using Python and SQL. The cumulative energy per production cycle per product/weight provides discrete values. This means that a value is stored in the database for each key figure and production cycle. The classification according to value added, on the other hand, provides time series that are divided into segments according to the above definition. These time series can in turn be evaluated for certain time windows (e.g., value-adding processes per month).

The required energy and production data were consolidated in the in-house SCADA system and the necessary calculations were performed using Python and SQL. The cumulative energy per production cycle per product/weight provides discrete values. This means that one value is stored in the database for each key figure and production cycle. The classification according to value creation, on the other hand, provides time series that are divided into segments according to the above definition. These time series can in turn be evaluated for certain time windows (e.g., value-adding processes per month).

Results

As part of the project, a comprehensive solution for digitizing the energy efficiency concept in an energy-intensive rolling mill was developed and implemented. The collected and processed information is made available via a specially developed dashboard in the company's own SCADA system (see Figure 2). At the beginning, the user selects the relevant time window and a specific calendar. This is followed by a detailed display of all batches processed during this period. After starting the analysis process, the system calculates various key figures that enable a differentiated analysis of energy consumption. The visualizations include, among other things, the energy consumption per batch and per material weight input. Particular emphasis is placed on the distinction between value-adding and non-value-adding energy consumption, which allows for precise identification of optimization potential within the production processes. In addition, the energy consumption of individual rolling processes is displayed in the form of a box plot, which allows irregularities and outliers in the energy requirements of individual production steps to be identified. The underlying time series for the various value-added contributions are analyzed for each selected time window, thus providing a structured overview of the course and efficiency of production.

With its clear and adaptable display, the dashboard supports targeted monitoring and comparison of energy consumption data for different batches and product variants. Noticeable deviations are

immediately visible, allowing both technical malfunctions and unusual energy consumption to be identified. These findings enable targeted fault diagnosis and also provide starting points for further measures to increase energy efficiency.

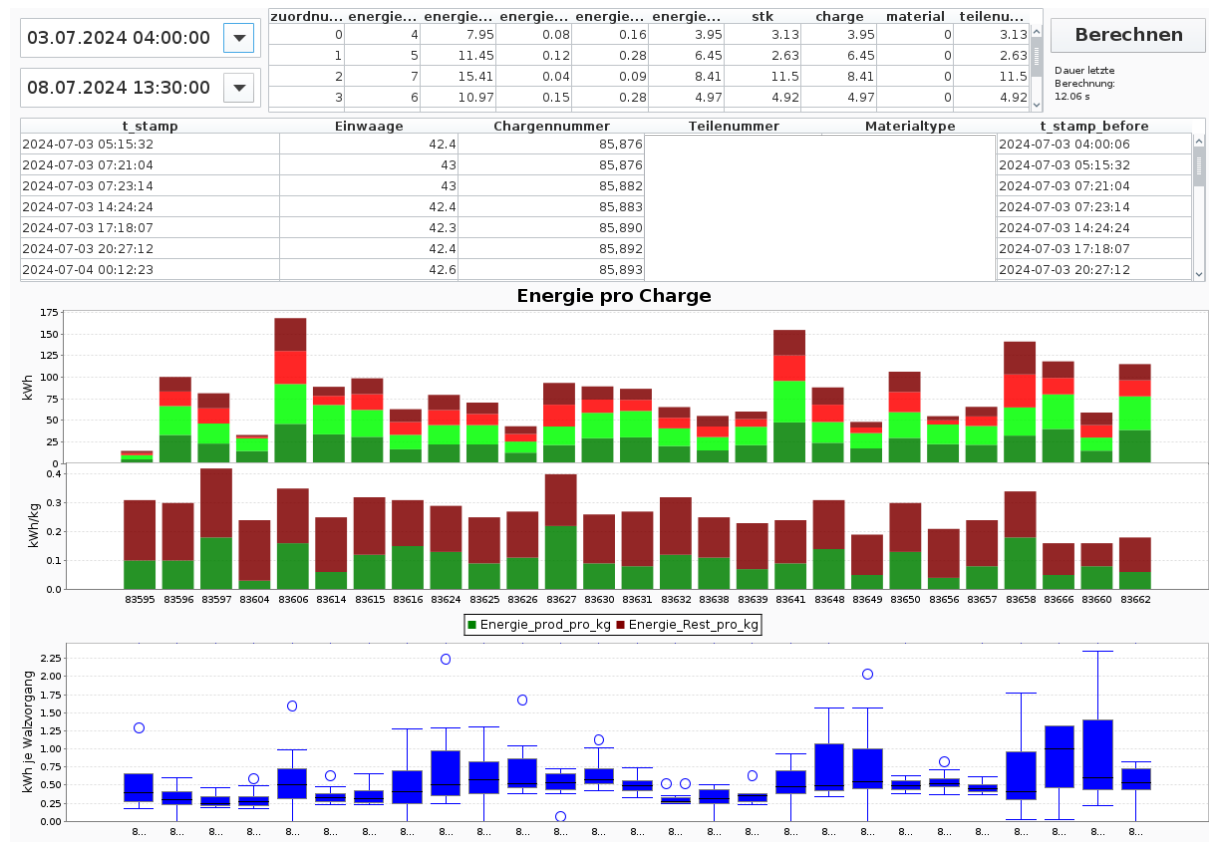


Abbildung 2. Dashboard

A central focus of the project was the universal applicability of the developed system architecture and methodology. Both the system itself and the chosen approaches for collecting and contextualizing the data were designed to be modular so that they can be transferred to other production lines or industrial sectors with comparatively little effort. The definition of key performance indicators and the segmentation of operating states can be flexibly adapted to different technical conditions and industry requirements. In addition, standardized interfaces enable smooth integration into other SCADA or operational management systems, thus ensuring broad compatibility of the developed solution. The visualization mechanisms and evaluation tools can also be adapted to new production processes, machine types, or reporting structures without fundamental changes. In particular, the methods for identifying and reporting anomalies in energy consumption prove to be universally applicable and can be directly connected to a wide variety of industrial applications. This holistic approach not only promotes optimization and error reduction in the rolling mill under consideration, but also creates the conditions for sustainable generalization and transferability to other industrial contexts, regardless of industry, product, or company size.

Summary

This article deals with an energy-intensive rolling mill used in the production of sealing materials. The central components are calendars consisting of two heated rollers that smooth and compact the raw material. Operation requires electrical energy and steam, the latter mostly from fossil sources, which increases the relevance of energy optimization. To save energy and improve the carbon footprint, the

existing energy monitoring and management system is being expanded to enable the contextualization of energy data with production data.

Specific key figures will be determined, e.g.:

- Energy consumption per material weight
- Energy consumption per product
- Energy consumption per operating state (direct, indirect, non-value-adding)
- Breakdown by energy types

The conceptual modularity and the key figures used can also be transferred to other production processes and machines with minimal effort. This means that both large industrial companies and smaller production facilities can benefit from the approaches and the system. Interfaces and visualization options are universally designed and can be flexibly adapted to different IT and production environments.

The analysis is performed by data processing with Python and SQL in the existing SCADA system. The results are visualized in a dashboard that displays energy consumption data per batch and process step – also broken down according to their contribution to value creation. The system helps with both energy optimization and error detection in production, for example by identifying outliers in the energy consumption of similar batches.

References

- [1] International Organization for Standardization, 23 07 2025. [Online]. Available: <https://www.iso.org/standard/69426.html>.
- [2] M. Hacksteiner, G. Fuchs und F. Bleicher, „Strategic energy management in mechanical series production: An industrial use-case,“ *Procedia Manufacturing*, 2019.

Casting process for sand molds with dynamic vacuum control ROBOCAST V

Edmundo Oliveira
Fill G.m.b.H
Edmundo.oliveira@fill.co.at

Sophie Binder
Fill G.m.b.H
Sophie.binder@fill.co.at

Kurzzusammenfassung

In diesem Beitrag wird eine innovative Simulationsstrategie für das ROBOCAST V-System im Rahmen einer TBI-Initiative vorgestellt. Mithilfe der Kombination von Finite Element Method (FEM) und Computational Fluid Dynamics (CFD) sowie zeitgesteuerten Widerstandsfiltern konnten dynamische Gießprozesse effizient und ressourcenschonend modelliert werden. Dies ermöglichte eine risikominimierte, datenbasierte Bewertung von Prozessflexibilität, Gussqualität und Systemeffizienz und zeigt den Nutzen digitaler Simulationen für die Optimierung industrieller Gießverfahren auf.

Keywords: Füllsimulation, dynamisches Netz, Filter-Trick

Abstract

This paper presents an innovative simulation strategy for the ROBOCAST V system as part of a TBI initiative. Using a combination of FEM (finite element method) and CFD (computational fluid dynamics) as well as time-controlled resistance filters, dynamic casting processes could be modeled efficiently and, in a resource-saving manner. This enabled a risk-minimized, data-based evaluation of process flexibility, casting quality, and system efficiency, demonstrating the benefits of digital simulations for optimizing industrial casting processes.

Keywords: Filling simulation, dynamic network, filtering trick

Introduction

In the rapidly evolving field of modern foundry technology, process efficiency and casting quality are increasingly becoming the focus of attention. A German foundry has recognized the transformative potential of digital simulation and initiated a TBI project to comprehensively evaluate the ROBOCAST V casting system prior to purchase. This approach enables the company to test the advantages and integration possibilities of the system in a risk-minimized, virtual environment—instead of immediate physical implementation.

Using advanced simulation tools, including the finite element method (FEM) and computational fluid dynamics (CFD), critical parameters of the ROBOCAST V dynamic casting process were examined in the context of specific production requirements. Digital models were used to gain insights into process flexibility, casting quality, and operational efficiency in order to support data-driven investment decisions.

This article describes the technical methods used and the challenges encountered during the simulation. It also highlights the added value of innovative digital strategies for mapping and optimizing complex casting processes. The following sections provide a more detailed explanation of the simulation techniques used, the special capabilities of the ROBOCAST V shown in Figure 1, and the practical implications for industrial foundries.



Figure 1. ROBOCAST V

Problem-solving approach

In collaboration with MAGMASOFT, a leading provider of casting simulation software, alternative strategies for state-of-the-art simulation were investigated. MAGMASOFT uses a method in sand casting simulation whereby a thin solid plate made of the same material is placed at the bottom of the sprue funnel. This plate gradually melts, allowing controlled flow into the gates and cavities. To replicate this principle, MAGMASOFT implemented a filter model that acts as a highly resistant barrier (similar to a wall) and can be switched on and off in a time-controlled manner. When switched on (“ON”), the filter prevents the flow of material; when switched off (“OFF”), the liquid metal can flow unhindered.

This well-known principle was used to simulate the movement of the ROBOCAST V casting pipe. The continuous geometry of the filling pipe was divided into several small segments (see Figure 2), with each segment being assigned its own time-based control function (see Figure 3).

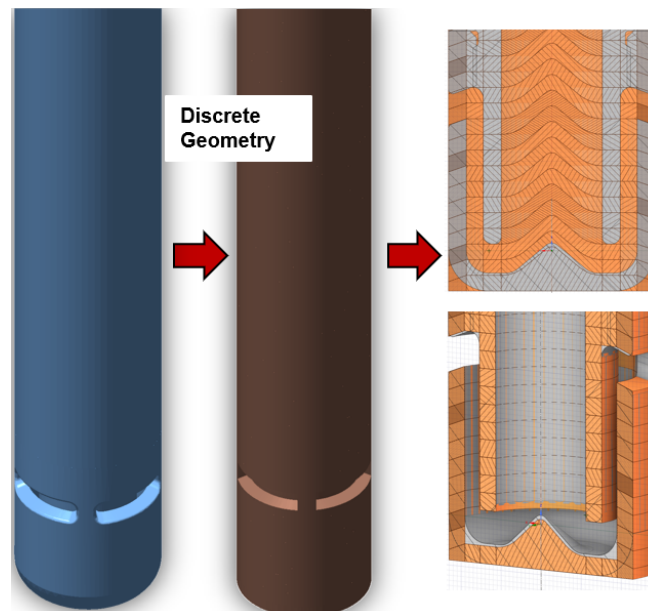


Figure 2. Discrete geometry

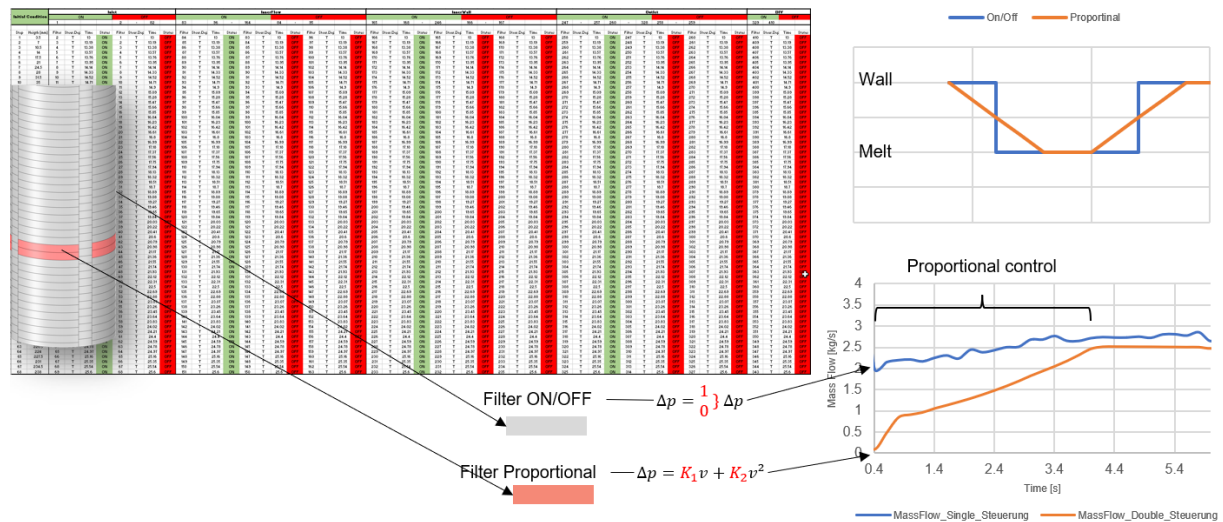


Figure 3. Control table (filter trick)

This method eliminated the need for dynamic mesh refinement (remeshing) while maintaining acceptable simulation times for complex casting geometries. MAGMASOFT does not allow pre-filled volumes as initial conditions, which means that the filling pipe must be completely filled during the filling simulation before the mold cavity can be filled. This posed an additional challenge: the filters at the outlet of the pipe (shown in pink in Figure 3) had to remain closed until the pipe was completely full. Abrupt opening led to unrealistic pressure peaks, as shown in Figure 4.

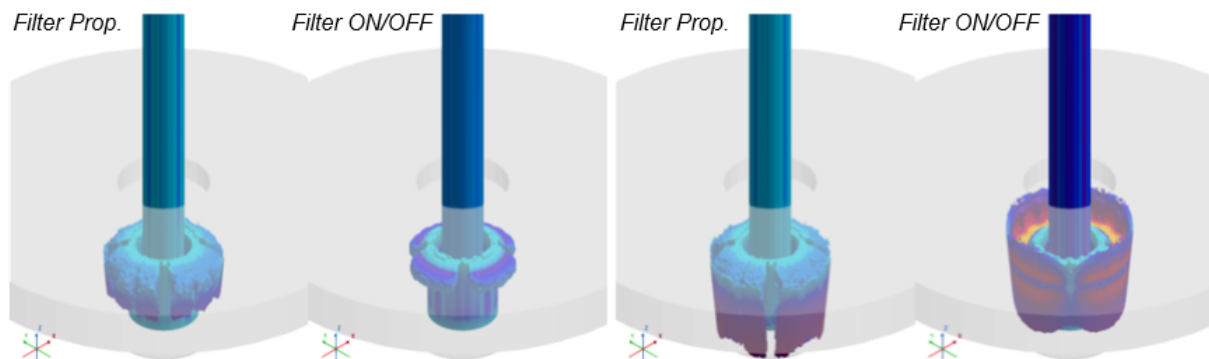


Figure 4. Overshoot due to pressure spike

To solve this problem, a progressive filter control was implemented. Reverse engineering was used to determine constants (K_1 , K_2) for the pressure loss equation of the filters, and a ramped adjustment was applied instead of a sudden change (0, 1) to simulate the gradual increase in mass flow observed in real processes (see Figure 4). The use of progressive filter control also improves the modelability of metal injection.

Results

As a result, simulation times suitable for industrial applications were achieved without compromising accuracy. This approach showed a high degree of consistency with real-world processes (see Figure 5).

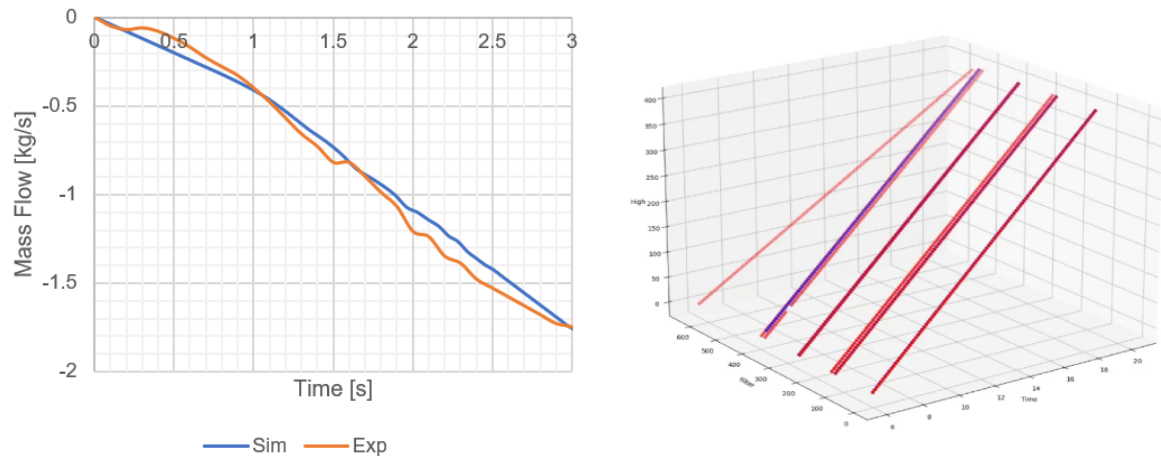


Figure 5. Comparison of mass flow and filter status

The filter control configuration takes only a few minutes and can be validated before the actual simulation. The control diagram (Figure 5) shows the ON/OFF status, the activation times, the filter ID, and the respective Z-axis position. Figure 6 shows the filling simulation using the dynamic filling pipe of the ROBOCAST V.

The model delivered such good results that MAGMASOFT developed a special module and integrated it into the new version 6.1. The module has been simplified and the user interface of the software has been made more intuitive. This means that any MAGMASOFT user who wants to simulate their process with ROBOCAST V can purchase this module as an additional function.

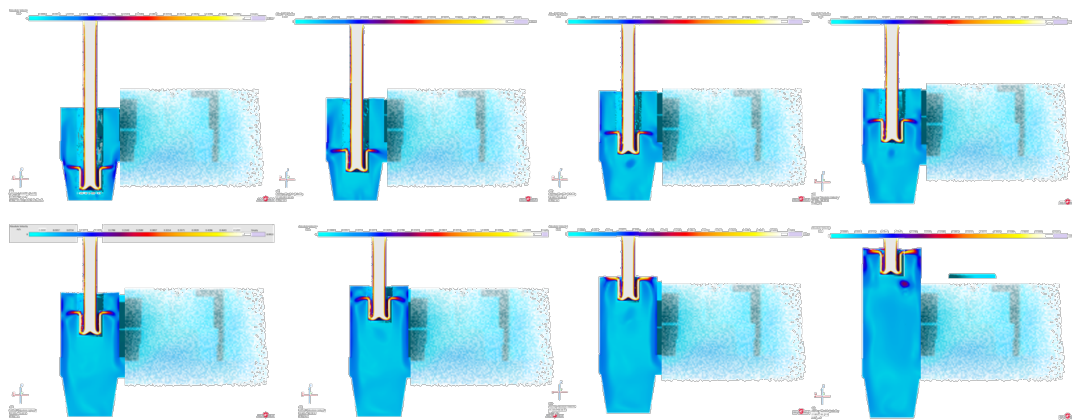


Figure 6. Filling simulation

Summary

This article presents a practical method for simulating complex aluminum casting geometries using the ROBOCAST V system and MAGMASOFT software. The progressive filter control developed replaces computationally intensive remeshing approaches and uses pressure loss constants determined by reverse engineering to realistically represent the gradual increase in mass flow. Challenges such as filling the filling pipe in front of the mold cavity and avoiding sudden pressure peaks were addressed. Visualization tools and control plots enable quick configuration and validation, while the simulation shows a high degree of consistency with practical processes. The integration of a dedicated ROBOCAST V module in MAGMASOFT version 6.1 simplifies application and increases accessibility, enabling industrial users to perform efficient and accurate process simulations that were not previously possible. All potential users (foundries, OEMs, research institutes, etc.) of a ROBOCAST V or similar casting system benefit from this development.

CFD simulation as a digital service for development and production

Judith Heim
Linz Center of Mechatronics GmbH
judith.heim@lcm.at

Markus Schörgenhumer
Linz Center of Mechatronics GmbH
markus.schoergenhummer@lcm.at

Martin Trinkl
Linz Center of Mechatronics GmbH
martin.trinkl@lcm.at

Johannes Klinglmayr
Linz Center of Mechatronics GmbH
johannes.klinglmayr@lcm.at

Alexander Dorninger
Linz Center of Mechatronics GmbH
alexander.dorninger@lcm.at

Daniel Polaschek
HAUSER GmbH
Daniel.Polaschek@hauser.com

Kurzzusammenfassung

Numerische Simulationen gewinnen in der industriellen Produktentwicklung zunehmend an Bedeutung. Ihre Anwendung erfordert bislang jedoch umfangreiche Fachkenntnisse und spezialisierte Software- und Hardware-Infrastruktur. Das stellt speziell für kleine und mittlere Unternehmen (KMU) eine Hürde dar. In diesem Beitrag wird ein cloudbasierter Webservice am Beispiel einer Strömungssimulation (CFD-Simulation) für Kühlmöbel vorgestellt. Er wurde entwickelt, um die Bereitstellung von Simulationen in einer Cloudumgebung via Web-Interface zu ermöglichen. Angewendet und getestet wurde der Webservice im Entwicklungs- und Optimierungsprozess von Kühlmöbeln durch Ingenieur:innen, die keine ausgewiesenen CFD-Simulationsexperten sind. Um dies ohne die Anschaffung von teuren Softwarelizenzen zu ermöglichen, wird dabei auf die Verwendung des Open-Source-Simulationstools OpenFOAM 9 gesetzt. Durch die Kombination aus automatisierter Geometrieerstellung, Vernetzung und einer vordefinierten, parametrisierten Simulation wird es auch Anwendern ohne einschlägige Simulationsexpertise ermöglicht, CFD-basierte Strömungsanalysen durchzuführen. Der Webservice unterstützt somit eine frühzeitige Integration von Simulation in den Entwicklungsprozess.

Keywords: CFD, OpenFOAM, SaaS, Webservice

Abstract

Numerical simulations are becoming increasingly important in industrial product development. Up to now, their application has traditionally required extensive expertise and specialized software and hardware infrastructure. This is a particular challenge for small and medium-sized enterprises (SMEs). This article presents a cloud-based web service using the example of a flow simulation (CFD simulation) for refrigerated display cases. It was developed to enable simulations to be provided in a cloud environment via a web interface. The web service was applied and tested in the development and optimization process of refrigerated display cases by engineers who are not proven CFD simulation experts. To make this possible without purchasing expensive software licenses, the open-source simulation tool OpenFOAM 9 is used. The combination of automated geometry creation, meshing, and a predefined, parameterized simulation enables users without relevant simulation expertise to perform CFD-based flow analyses. The web service thereby supports the early integration of simulation into the development process.

Keywords: CFD, OpenFOAM, SaaS, Webservice

Introduction

When developing refrigerated display cases, it is essential to ensure during the design process that the construction achieves the most uniform temperature distribution possible inside the interior and low

energy consumption. In the case of open refrigerated display cases in particular, the interaction between air flow and heat transfer inside and outside the product compartment is very complex, as a cool air curtain must separate the product compartment from the warm environment.

Conventionally, refrigerated display cabinets are developed in an iterative process involving design, prototyping, and production. This involves high prototype and testing costs, including comprehensive standard-compliant measurements and tests. The use of digital development methods based on numerical simulations can significantly support this process. Furthermore, this opens up new opportunities for further targeted optimization of the products, so that energy can be saved during the subsequent operation of the refrigerated display cases, thus making an important contribution to sustainability. In addition, resources, effort, time, and costs in product development can be reduced.

The difficulty lies in the fact that the necessary flow simulation method (CFD simulation) is, on the one hand, very complex and therefore difficult for non-experts to access and, on the other hand, involves a great deal of effort and, as a rule, high costs for software and hardware. In order to make these digital simulation tools accessible to users without extensive simulation expertise for a specific area of application, a robust, parameterized, and automated simulation model is required. This is implemented in a cloud environment and made available to users via an easily accessible and intuitive web interface. In order to create good usability and functionality for users, a compromise must be made between general validity, level of detail, and accuracy. In addition, the computational effort of the models must also be taken into account with regard to accuracy. Furthermore, it is essential to ensure the robustness of the modeling and simulation workflow in order to enable implementation of the web service that can also be reliably used by non-experts.

Against this backdrop, research and development is increasingly focusing on automated, parameterizable, and cloud-based simulation solutions that aim to make simulation applications more accessible. On the one hand, this shifts the computing load to the cloud, and on the other hand, automated networking algorithms, automated simulation workflows, and web-based user interfaces enable low-threshold access to simulation. This is gradually leading to a paradigm shift from simulation-driven product development carried out by experts to more accessible, design-integrated simulation support that is also available to non-simulation experts.

This phenomenon can also be seen among large software manufacturers, who are increasingly relying on cloud- and web-based services with semi-automated workflows. For example, ANSYS [2], Siemens [3], and AVL [4] now offer “Simulation as a Service” (SaaS) via web- and cloud-based applications. Other software manufacturers such as SimScale [5] are focusing not only on operating simulations via a web interface and computing them in the cloud, but also on using AI-supported automated workflows to support engineers from the creation of the simulation to the analysis of the results.

The aim of this article is to use the example of the development of refrigerated display cabinets to present a functional prototype for a web service that can also be used by non-simulation experts in the product development and optimization process. Open-source simulation tools are used to make this possible without the need to purchase expensive software licenses.

Solution approach

CFD-Modell

To enable efficient testing of the functional prototype, a greatly simplified test setting was used for the CFD simulation model. This allows the functionality of the prototype to be tested in a reasonable amount of time using a wide variety of scenarios with a reasonable amount of computing power. The open source software OpenFOAM 9 [1] is used for the CFD simulation. OpenFOAM 9 integrates the BlockMesh and SnappyHexMesh meshers, which are used to mesh the geometry. Both the meshing and the simulation can be performed in the functional prototype either on one CPU core in “single” mode or on four CPU cores in “parallel” mode.

Geometrie

The geometry for the test setting used in the functional prototype was based on the geometry of an open refrigerated display case located in a test room. For the sake of simplicity, the shelves of the refrigerated display case were omitted from the geometry and the structure of the rear wall was simplified. Figure 1 shows the simplified geometry of the refrigerated display case in the test room. The refrigerated cabinet geometry is stored in the functional prototype as an *.stl file. Only the overall size of the refrigerated cabinet can be varied in the functional prototype. The stl file containing the geometry of the refrigerated cabinet is scaled in all three coordinate directions by a specified factor, which can be set as a parameter in the web interface.

The size of the cuboid test room can be freely selected by the user of the framework by defining the corner points (X Min, Y Min, Z Min) and (X Max, Y Max, Z Max) along the main diagonal of the cuboid.

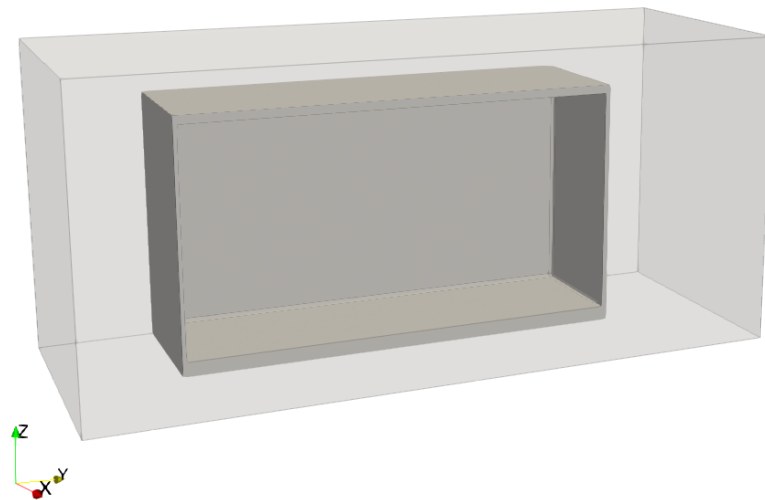


Figure 1. Simplified geometry used in the functional prototype.

Mesh generation

BlockMesh is used to create a cuboid background mesh for the test space. The resolution of this mesh in each coordinate direction can be specified by the user using the parameters X Discr, Y Discr, and Z Discr. These parameters determine how many cells are generated when meshing the cuboid in the respective spatial direction.

After creating the background mesh, SnappyHexMesh is used to cut out the geometry of the refrigerated display case from the background mesh of the test space. The mesh is approximated to the surface of the *.stl geometry by iteratively refining the background mesh by halving the cells near the surface of the *.stl geometry. Following this refinement, the cells that lie outside the fluid volume are removed. This results in a mesh that maps the surfaces of the *.stl geometry in a very stepped manner, as the cells along the edges are cuboid. Therefore, the cell nodes that lie along the surface of the *.stl geometry are then moved to the surface in order to smooth the stepped surfaces and better adapt them to the surface.

In order to be able to sufficiently resolve even large velocity gradients near the walls in the simulation, additional layers with hexahedral cells, known as inflation layers, are inserted at the wall surfaces. These inserted cells are very thin directly at the wall, with the layer thickness then increasing exponentially from the walls into the interior of the simulation area [6]. The settings for SnappyHexMesh are fixed in the function prototype and cannot be changed by the user, as this would require expertise in meshing geometries for CFD simulation in order to specifically influence and optimize the mesh quality.

Simulation model

The simulation in this functional prototype is used to test the functionality of the web interface and the “simulation-on-demand” concept for non-simulation experts. To enable this efficiently, the complexity of the simulation has been greatly reduced. A stationary, isothermal, and incompressible flow is assumed. The user of the functional prototype can specify the pressure at the outlet from the simulation area or ambient pressure and the vectorial flow velocity at the inlet to the simulation area as boundary conditions. In addition, the user can use the End time parameter to set how many iterations are to be calculated in this steady-state simulation. The remaining boundary conditions, as well as the density and viscosity of the air and all other numerical settings, are fixed in the functional prototype.

If you select the “Parallel” mode for running the simulation, the simulation will be executed in parallel on 4 CPU cores in the function prototype. This can speed up the simulation for networks with many cells. If the network has very few cells, a parallelized simulation can be counterproductive because the speed-up achieved through parallelization is less than the additional overhead caused by the extra communication between the CPU cores. As a guideline, parallelization is useful in this application if there are more than 25,000 cells per processor core.

Display of simulation results

The functional prototype features automated display of simulation results. This involves generating *.vtk files, which form the database for displaying the mesh and simulation results in the viewer. The results are processed automatically after the simulation. Only surfaces can be displayed in the web viewer, not volumes. Therefore, the *.vtk files automatically created with OpenFOAM must first be converted to *.vtp files in order to load and view the mesh and simulation results in the web viewer. All boundary surfaces in the simulation model are available here, as well as pressure and velocity on selected cut surfaces defined in the simulation model by the geometry.

To enable the simulation to be deployed in a cloud environment via a web interface, a Python module was developed that allows users to set input parameters in the network and simulation. The Python module for configuring user parameters can only be accessed from external services via an application programming interface (API), which also enables communication via a web interface. In this way, the simulation framework can be provided as a web service and integrated into external applications or systems.

In conjunction with a containerized OpenFOAM, the API provides the basis for decentralized simulation in a cloud environment without the need for hardware and software resources on the user's side. Via a web interface, users of the web service can use the API to change the available parameters via a graphical user interface, start the networking and simulation, and display the simulation results.

The functional prototype presented in this article was implemented in the Amazon Web Services [7] cloud environment using Amazon Elastic Container Service (ECS) [8] on EC2 [9] and made available to users.

Results

Figure 2 shows the input masks for geometry creation, meshing, and simulation. Figure 3 shows a contour plot of the velocity magnitude on the intersection plane normal to the y-direction through the center of the geometry as an example of how the simulation results are displayed.

By automating meshing and simulation creation using API and Python modules, end users of the web service do not need in-depth technical or numerical knowledge to use the simulation as a tool in product development. Since a cloud environment is used to perform the simulation, it can be used anytime, anywhere. Users do not need to purchase special, powerful hardware or install software, as the

simulation is designed as a web service and can be operated via a web interface in the browser.

Base Mesh

X Min:

X Max:

Y Min:

Y Max:

Z Min:

Z Max:

X Discr:

Y Discr:

Z Discr:

Resize STL

X Min:

Run Mesh

Mode:

Simulation Settings

Start Time:

End Time:

Initial Conditions

Inlet Velocity:

Ambient Pressure:

Run Simulation

Mode:

Figure 2. Input mask of the web interface for automated networking (left) and for automated simulation (right).

In the future, it is also conceivable to add an optimizer to the web service and thus, for example, perform automated geometry optimizations. To do this, it would also be necessary to integrate an interface to open-source CAD software so that the geometry can be generated in a parameterized manner and thus also modified easily and automatically. This would make numerically supported optimization accessible even to non-simulation experts and provide significant support in the design process for new products.

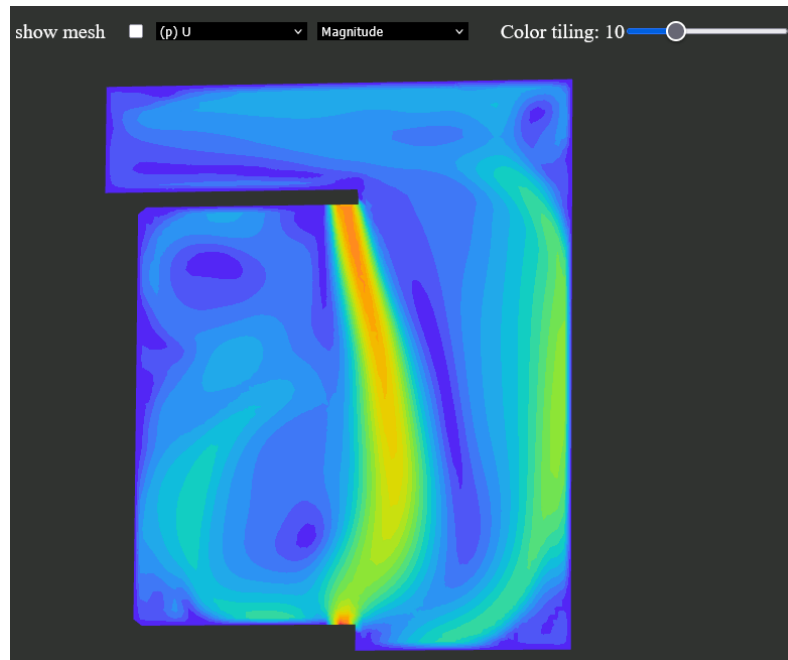


Figure 3. Exemplary contour plot of the velocity magnitude.

Furthermore, the developed framework can be adapted so that it can also be used for simulations from other physical domains, e.g., FE simulations, with both open-source tools and commercial software. The web service is therefore very versatile and can also be adapted to individual needs and used beneficially in other industries and companies in the future.

Summary

A functional prototype for the automated remote calculation of a suitably simplified CFD simulation model was created and tested using the example of refrigerated display cabinet development. This demonstrated that a good compromise can be found between usability and functionality or level of detail for the areas of application and requirements. During practical testing of the web service by non-simulation experts, its accessibility and ease of use were rated very positively. It was also emphasized that the web service provides a good basis for the low-threshold integration of CFD simulation into the development process of refrigerated display cases. Several points were also identified where there is still room for expansion and improvement, both at the GUI level and in terms of technical functionality. The isothermal simulation of the functional prototype can be expanded to a thermal simulation. The integration of an interface to open-source CAD software is also desirable so that different refrigerated display cabinet geometries can be generated in a parameterized manner and varied easily.

Furthermore, the framework can also be used for simulations from other physical domains, such as FE simulations, by making various adjustments. In addition to open-source tools, commercial software can also be integrated into the framework.

In this way, a web service for numerical simulation can be set up that can be used in a low-threshold manner to support the design and development process of products. This opens up new potential for saving time and costs.

References

- [1] H. Weller, C. Greenshields, und W. Bainbridge, *OpenFOAM 9*. (20. Juli 2021). C++. The OpenFOAM Foundation Ltd, London. [Online]. Verfügbar unter: <https://openfoam.org/release/9/>
- [2] „Simulation as a Service: Ansys Lizenzen & Hardware aus der Cloud“. Zugegriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://www.cadferm.net/at/de/unsere-loesungen/simulation-as-a-service.html>

- [3] „Simcenter X - HPC and remote desktop“, Siemens Digital Industries Software. Zugriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://plm.sw.siemens.com/de-DE/simcenter/simcenter-x/hpc/>
- [4] „Simulation as a Service | AVL“. Zugriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://www.avl.com/en/simulation-solutions/simulation-a-service>
- [5] „Simulation Software | Engineering in the Cloud“, SimScale. Zugriffen: 23. Juli 2025. [Online]. Verfügbar unter: <https://www.simscale.com/>
- [6] C. Greenshields, „OpenFOAM v9 User Guide“, CFD Direct. Zugriffen: 22. Juli 2025. [Online]. Verfügbar unter: <https://doc.cfd.direct/openfoam/user-guide-v9/snappyhexmesh/>
- [7] „Amazon Web Services AWS – Server Hosting & Cloud Services“, Amazon Web Services, Inc. Zugriffen: 25. Juli 2025. [Online]. Verfügbar unter: <https://aws.amazon.com/de/>
- [8] „AWS | Amazon EC2 Container & Konfigurationsmanagement“, Amazon Web Services, Inc. Zugriffen: 22. Juli 2025. [Online]. Verfügbar unter: <https://aws.amazon.com/de/ecs/>
- [9] „Amazon EC2 – Cloud-Datenverarbeitungskapazität – AWS“, Amazon Web Services, Inc. Zugriffen: 22. Juli 2025. [Online]. Verfügbar unter: <https://aws.amazon.com/de/ec2/>

Intelligent chatbots for querying company knowledge- ALLEX

Pichlmann Lukas
LKR Leichtmetallkompetenzzentrum
Ranshofen GmbH
Lukas.Pichlmann@ait.ac.at

Kurzzusammenfassung

Die Implementierung eines intelligenten Chatbots Advanced Language Learning for Engineering Excellence (ALLEX) zur Abfrage von Unternehmenswissen stellt eine nutzerfreundliche Lösung zur Verbesserung der Informationszugänglichkeit dar. Dieser Beitrag beschreibt die Entwicklung eines Chatbots basierend auf Retrieval Augmented Generation, welcher Nutzeranfragen zu firmeninternen Dokumenten mittels lokaler Large Language Models (LLMs) beantwortet. Die Lösung nutzt Open Source Komponenten und kombiniert semantische Suche mit lokalen LLMs. Das System ermöglicht es Mitarbeitern und Mitarbeiterinnen, effizient auf historische Daten und Dokumente zuzugreifen und dabei von intelligenter Kontextualisierung zu profitieren.

Keywords: RAG, Chatbot, Large Language Models

Abstract

The implementation of an intelligent chatbot for querying company knowledge represents a user-friendly solution for improving information accessibility. This contribution describes the development of a chatbot based on Retrieval Augmented Generation, which answers user queries about internal company documents using local Large Language Models (LLMs). The solution utilizes open source components and combines semantic search with local LLMs. The system enables employees to efficiently access historical data and documents while benefiting from intelligent contextualization.

Keywords: RAG, Chatbot, Large Language Models

Introduction

Over the years, companies accumulate extensive data sets of documentation and empirical knowledge. This information is often difficult to access or locate and is therefore not used optimally. Employees benefit from a user-friendly way to access documented cases and processes within the company, identify similar situations, and learn from historical solutions.

Retrieval Augmented Generation (RAG) combines the strengths of information retrieval and text generation, enabling specific domain knowledge to be integrated into large language models (LLMs) without the need for time-consuming fine-tuning processes.

The approach implemented here can be applied across industries and transferred to various corporate contexts to optimize the retrieval of internal company knowledge and provide employees with precise, context-related answers to their questions.

Implementing an RAG system requires various complex technologies, the development of which would go far beyond the scope of this article. Therefore, proven open-source (OS) components were used, which benefit from continuous development and the collective knowledge of the community while ensuring comparatively high quality and stability. In addition, care was taken to use small and resource-efficient LLMs in order to minimize hardware requirements and make the system accessible with as little investment as possible.

Problem-solving approach / Procedure

System architecture

Three different open source projects were selected for the implementation of ALLEX:

- Hosting of local LLMs: Ollama
- Hosting of embedding and reranking models for RAG: XInference
- Platform with RAG support, web interface, and (low-code) functionality for programming chatbot logic: Dify

All components can be hosted locally, offline, and containerized using Docker and offer a permissive license, which was one of the most important criteria in the technology selection process.

Ollama

Ollama is a tool developed to facilitate the deployment and local application of large language models (LLMs) [1]. Ollama also offers a comprehensive model library with a wide selection of state-of-the-art open weight LLMs that can be used with minimal effort.

Ollama was selected because it is a popular tool with a very large open-source community, which ensures a significant degree of stability. In addition, it is easy to install and use, allows various LLMs to be run locally without complex configuration, and has an active developer community that regularly provides updates and improvements.

XInference and semantic search

Similar to Ollama, XInference provides a way to host LLMs locally, as well as other AI models [2]. Since Ollama did not offer the option to host so-called reranking models at the time ALLEX was developed, XInference was selected as an additional component of ALLEX. In ALLEX, XInference is used to provide the necessary models for subsequent semantic search.

To perform semantic search, embedding models are required. These are specialized in representing natural text and its components as (high-dimensional) numerical vectors that represent the semantic context of specific text passages. This allows information, such as internal company documents and the text they contain, to be mapped using machine-readable numerical values. This numerical data can then be stored in a vector database, making it available for semantic search. Such embedding models are also used later to convert incoming user queries into numerical vectors. Using these vectors and a predefined distance metric (e.g., cosine similarity), the vector database can then be searched for similar entries (i.e., vectors with a small distance to the vector resulting from the query).

Entries from the database are ranked according to relevance using distance metrics. To improve the accuracy of this ranking, optional reranking models can be used. These models evaluate the candidates found by the initial vector search in greater detail and reorder them based on an (improved) assessment of their relevance to the user query. Reranking models typically work with cross-attention mechanisms, which can capture the semantic relationship between the query and the documents more precisely than pure vector distance.

XInference offers a comprehensive selection of open weights embedding and reranking models, which are used in ALLEX for the implementation of retrieval augmented generation.

Dify

Ollama and XInference ensure the local deployment of all necessary AI models for the system. However, the main work of this paper lies in the logic and programming of the chatbot, which is a combination of

LLM, embedding and reranking models, document database, and program logic. Dify was selected to create this program logic and configure the database [3].

Dify is a modular OS project that specializes in the development of LLM applications. It offers easy integration with a wide range of hosting services, including Ollama and XInference, and enables simple configuration of a vector database and the development of complex LLM workflows using a low-code platform.

Figure 1 shows a minimal example of a simple chat flow in Dify's low-code environment, where complex logic can be programmed between user queries (start node) and returned responses (answer node), either using the nodes and building blocks preconfigured in Dify or with your own source code, which can be executed in so-called code nodes using the Python programming language.

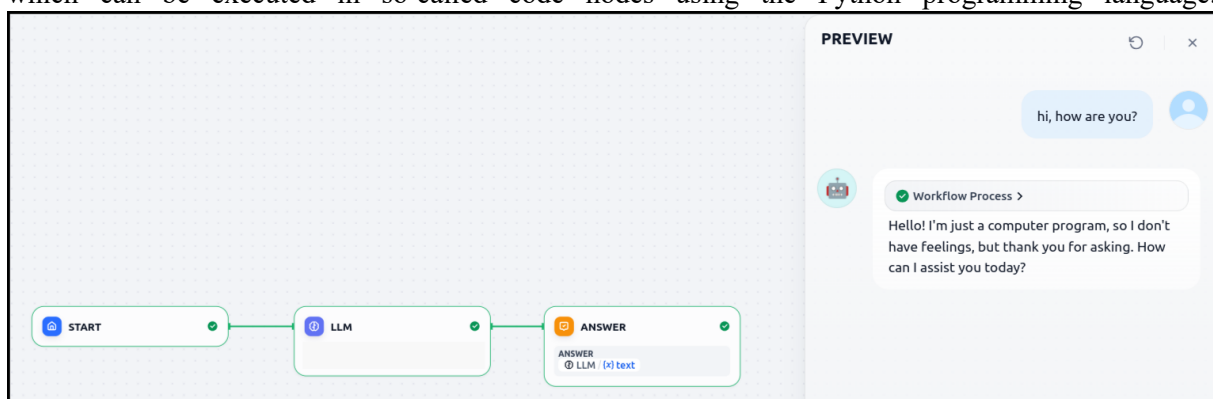


Figure 1. Example of a simple process in Dify's low-code environment, consisting of a start node, an LLM node, and an answer node for simple chat interaction with an LLM

These chat flows can be used to map complex workflows, which are then made available to users via a chat interface. Dify also offers the option of displaying external content used in the response via RAG to the user at the end of the response with the help of citations. This feature makes it possible to check the accuracy of the responses generated, thereby improving the transparency of the chatbot.

Dify also offers user-friendly database integration for using RAG, with a Weaviate vector database used by default.

Program logic

Figure 2 shows the chat flow developed during the course of the project, which was created using Dify (top), as well as a schematic representation of the most important functions it contains (bottom).

To create this chat flow, not only were the standard building blocks included in Dify used, but also self-developed nodes that had to be added and programmed in the source code of the Dify project: so-called “context” nodes. These allow the available context of each LLM agent (or each LLM node) in the chat flow to be dynamically adjusted and modified during the execution of the chat flow depending on the situation and decision path. This makes it possible, for example, to remove the document content that was made available to the LLM for answering via RAG from the context after generating a response to a query. This saves tokens in the context window, giving an LLM agent a longer “memory” (i.e., more past chat messages in context).

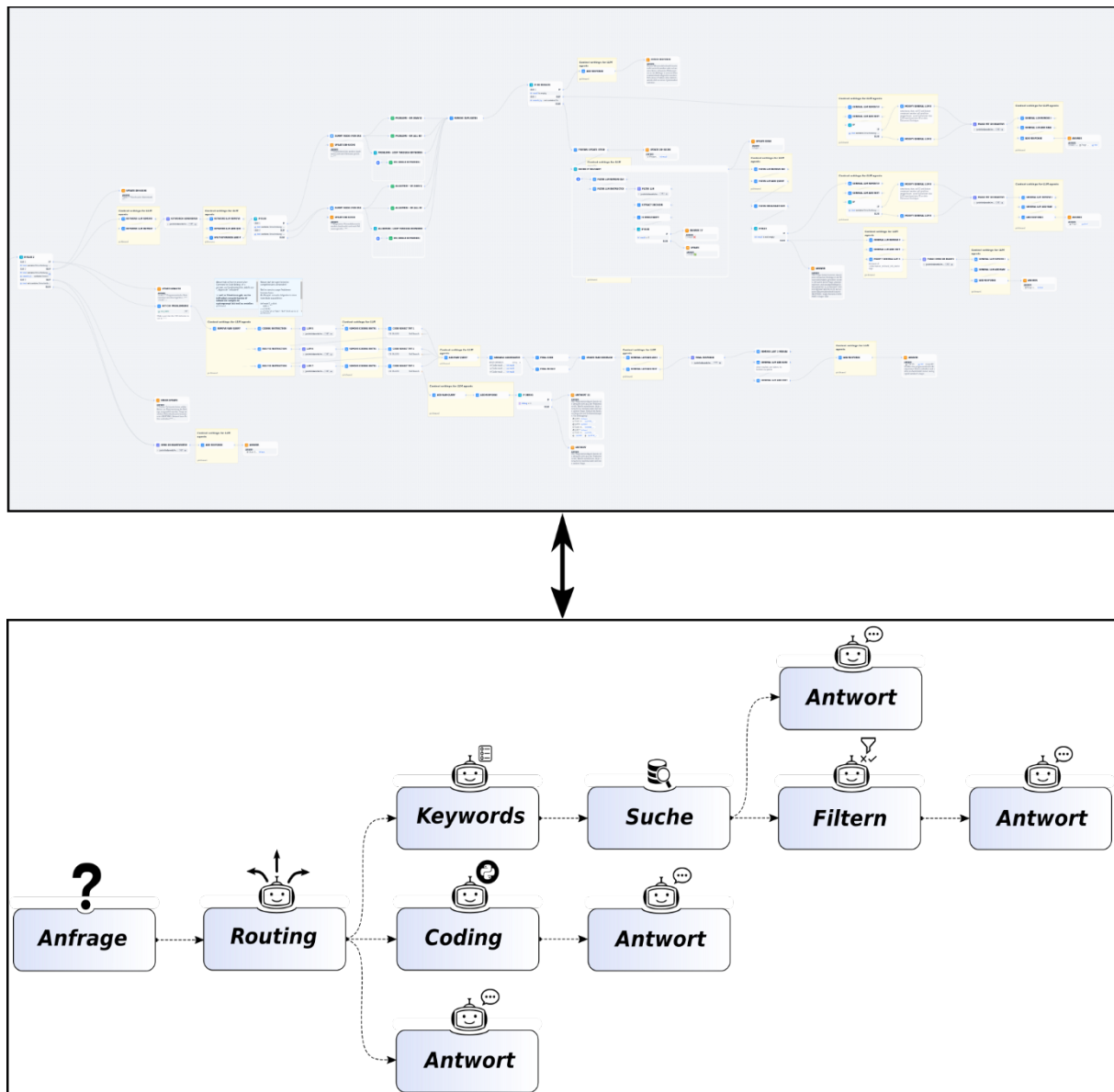


Figure 2. Screenshot of the workflow created in Dify (top) and corresponding schematic representation of the programmed logic (bottom)

The program sequence consists of various LLM agents with different tasks. A system prompt was created for each of the LLM agents involved, which explains the respective task in detail with a variety of examples. Each LLM agent also has its own chat history, which is dynamically adjusted during execution. A total of 5 different types of LLM agents are used:

- The “routing” agent decides which of the 3 possible paths in Figure 2 should be chosen for the current request and how it should be answered:
 - By answering directly without further context (lower path)
 - With the help of a coding agent that has access to a structured table of historical problem cases and can thus perform simple programmatic analyses (middle path)
 - Using RAG, whereby the query is answered by an LLM with the help of information from the database (upper path) Der “Antwort”-Agent ist dafür zuständig, die finale Antwort zu generieren, die dem Nutzer weitergegeben wird. Dieser wird in jedem der 3 Pfade verwendet.
- Der “Coding”-Agent ist dafür zuständig, Python-Code für einfach programmatische Analysen der Tabelle aus historischen Problemfällen zu generieren.

- The “Keyword” agent is responsible for generating various keywords from the user query, which are used together with the query for semantic search.
- The “Filter” agent is used optionally (depending on the user configuration) after the semantic search to decide for each of the selected entries whether the respective entry actually helps to answer the query or can be removed from the context. This helps to avoid filling the limited context of the LLM with irrelevant information.

Here is an example of the resulting workflow: The user asks the question “Have there ever been problems with incorrectly dimensioned components and what was done about it?” in the chat interface. The “routing” agent decides that semantic search (RAG) is required for this user query. The “keyword” agent receives the current query (as well as a limited chat history) and generates a handful of keywords and key phrases. Using the original query and the generated keywords, the vector database is searched for relevant entries, and the top N entries whose relevance score exceeds a certain threshold are made available (after optional filtering of irrelevant entries by the “filter” agent) to the “response” agent for generating a message, which is then output to the user, including the sources used.

If a query is made that requires statistical analysis or evaluation of the available structured table of past problem cases in the company (for example, “How many problem reports were there in year X and what were the average costs incurred?”), the “routing” agent should select the path to programmatic analysis. The code then generated by the “coding” agent is executed in a Python sandbox, and if successful, the result is passed to the ‘response’ agent to generate a response based on the evaluation. If the execution is not successful, the “coding” agent is allowed to correct the program code. To do this, the error message generated during execution is saved and made available to the agent for correction. The agent has a maximum of three attempts to generate valid Python code before the chat flow is terminated and a corresponding error message is displayed to the user.

All of the agents mentioned use the same underlying LLM and are executed sequentially, only with different contexts. This ensures that a single model located in the GPU memory can be used to operate the chatbot, which reduces hardware requirements.

Evaluation

A comprehensive questionnaire was developed to evaluate the quality of the system's responses and, based on this, to adjust the program logic, models used, and parameters. In it, ALPEX defined a wide variety of possible user queries as well as expected responses and relevant documents needed to answer them. This made it possible to objectively record how many of the relevant text excerpts and documents were actually found using semantic search when answering the query. In addition, an independent (local) LLM was used to evaluate how well the chatbot's response matched the expected response from the questionnaire (score from 1 to 5).

Since Dify enables programmatic execution of the chatbot via its own API, this evaluation process could be fully automated. To do this, a Python script was created that iterates through the questions and expected answers/documents in the question catalog, executes the chatbot for each question using the Dify API, and evaluates the resulting answers, including quotes, with the documents used. Metrics such as the number of relevant documents found and the evaluation of the match (by an LLM) with the predefined expected answer were then stored for each question and finally aggregated to quantify the performance of the system.

Based on this score, the program logic was revised, models were adjusted, and parameters in the semantic search were optimized.

The following models were ultimately selected for the prototype created:

- LLM: Mistral Small 3.1 24B (Q6) [4] [5]
- Embedding model: Multilingual e5 Large [6]
- Reranking model: Jina Reranker v2 [7]

The models mentioned showed promising performance and also represent a resource-efficient way to operate the system. With a comparatively low GPU VRAM requirement of around 30GB for LLM applications, the software can already be used performantly with commercially available consumer hardware such as the NVIDIA 5090 at the time of writing.

Results

The system offers users the opportunity to communicate with a chatbot using natural language, which dynamically decides how and with which tools to respond to user queries. Information from a database can be used to answer the query, or simple programs for data evaluation can be created and executed if necessary. The focus of the system is on being able to operate all components completely offline in order to ensure high data security. In addition, care was taken to use LLMs that are as small and resource-efficient as possible so that the system can also be operated with comparatively inexpensive hardware. The solution thus offers an efficient way to integrate modern technologies such as RAG into smaller companies without transferring internal data to cloud providers.

The system offers users the opportunity to communicate with a chatbot using natural language, which dynamically decides how and with which tools to respond to user queries. Information from a database can be used to answer the query, or simple programs for data evaluation can be created and executed if necessary. The focus of the system is on being able to operate all components completely offline in order to ensure high data security. In addition, care was taken to use LLMs that are as small and resource-efficient as possible so that the system can also be operated with comparatively inexpensive

Summary

This article demonstrates the practical implementation of a local RAG system for querying corporate knowledge using exclusively open-source components. The developed solution combines three established technologies (Ollama, XInference, and Dify) into a functional system that can be operated completely offline.

With the help of these OS tools, a multi-agent system with database access was developed in which various specialized LLM agents, based on the same basic model, work sequentially to answer user questions. This enables efficient use of resources while maintaining high functionality.

Using the low-code functionality in Dify, a workflow was created and iteratively adapted that enables user queries to be answered with the help of semantic search and an autonomous coding agent supported by information from the database. Automated evaluation using a questionnaire and LLM-based assessment proved to be an effective method for evaluating system performance and optimizing system and model parameters.

The solution shows that even with relatively modest hardware requirements (approx. 30GB GPU VRAM), an LLM application can be implemented that can improve the findability of relevant internal information. Such a system is particularly relevant for small and medium-sized enterprises, as it represents a cost-effective alternative to cloud-based enterprise solutions. Companies can retain complete control over their data while still benefiting from modern AI technologies. Furthermore, the exclusive use of open-source components means that there are no license costs or vendor lock-in effects.

The article thus provides a practical guide for companies that want to better tap into their internal knowledge resources using modern AI technologies without having to compromise on data security or cost efficiency.

References

- [1] „Ollama,“ [Online]. Available: <https://ollama.com/>. [Zugriff am 28 07 2025].
- [2] L. Weizheng, X. Lingfeng, Z. Feng, Q. Xuye, und C. Yueguo, "Xinference: Making Large Model Serving Easy". In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Seite 291–300, Miami, Florida, USA. Association for Computational Linguistics.
- [3] LangGenius, „Dify - GitHub,“ [Online]. Available: <https://github.com/langgenius/dify>. [Zugriff am 28 07 2025].
- [4] „Hugging Face - mistralai/Mistral-Small-3.1-24B-Instruct-2503,“ [Online]. Available: <https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503>. [Zugriff am 28 07 2025].
- [5] „Hugging Face - bartowski/mistralai_Mistral-Small-3.1-24B-Instruct-2503-GGUF,“ [Online]. Available: https://huggingface.co/bartowski/mistralai_Mistral-Small-3.1-24B-Instruct-2503-GGUF. [Zugriff am 28 07 2025].
- [6] „Hugging Face - intfloat/multilingual-e5-large,“ [Online]. Available: <https://huggingface.co/intfloat/multilingual-e5-large>. [Zugriff am 28 07 2025].
- [7] „Hugging Face - jinaai/jina-reranker-v2-base-multilingual,“ [Online]. Available: <https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual>. [Zugriff am 28 07 2025].

Optimization of forming processes through AI-based monitoring and control

Denk Michael
LKR Leichtmetallkompetenzzentrum
Ranshofen GmbH
Michael.Denk@ait.ac.at

Pichlmann Lukas
LKR Leichtmetallkompetenzzentrum
Ranshofen GmbH
Lukas.Pichlmann@ait.ac.at

Kurzzusammenfassung

Dieser Beitrag präsentiert ein System zur Echtzeitüberwachung von Zugproben mittels klassischer Computer-Vision-Methoden als Grundlage für eine adaptive Stromregelung im Umformprozess. Ziel ist die dezidierte Auslösung des elektroplastischen Effekts (EPE) durch eine dynamische Anpassung der Stromimpulse in Abhängigkeit der beobachteten Querschnittsveränderung der Probe. Die Geometrieanalyse erfolgt über eine Kombination aus Canny-Kantendetektion und RANSAC-Linienregression zur robusten Bestimmung des Probenabstands. Die gewonnenen Messwerte und davon abgeleiteten Regelparameter dienen unmittelbar als Eingangsgröße für die Steuerung der Hochstromanlage. Das System ermöglicht damit eine prozessnahe, sensorgestützte Steuerung und stellt eine innovative Lösung für die Echtzeitoptimierung industrieller Umformprozesse dar.

Keywords: Computer Vision, Zugprobe, Echtzeitüberwachung

Abstract

This contribution presents a system for real-time monitoring of tensile specimens using classical computer vision methods as a foundation for adaptive current control in forming processes. The objective is the targeted activation of the electroplastic effect (EPE) through dynamic adjustment of current pulses based on the observed cross-sectional changes of the specimen. The geometry analysis is performed through a combination of Canny edge detection and RANSAC line regression for robust determination of specimen spacing. The acquired measurement values and derived control parameters serve directly as input variables for controlling the high-current system. The system thus enables process-oriented, sensor-supported control and represents an innovative solution for real-time optimization of industrial forming processes.

Keywords: Computer Vision, Tensile Specimen, Real-time Monitoring

Introduction

The targeted triggering of the electroplastic effect (EPE) in metallic materials opens up new possibilities for reducing forming forces and increasing ductility. According to current research, the effect is highly dependent on the geometry, the material condition, and the timing of the high-current control. Real-time control of the current profile based on the actual mechanical state of the sample has not yet been achieved.

Previous methods have worked with rigid current profiles without feedback on the actual plastic thinning that occurs. The aim of this paper was to develop a system that allows adaptive control of the current pulses via image-based real-time recording of the sample geometry. This establishes an intelligent closed-loop control system that can be used directly in industrial forming processes.

Such an image-based control principle has the potential to be transferred to various forming processes and offers a cost-effective alternative to complex sensor technology, especially for small and medium-sized enterprises.

Problem-solving approach / Procedure

The developed system consists of a GUI application based on the following main components:

- Data acquisition: FastAPI-based interface for camera integration
- Camera calibration: Correction of image distortions
- Parameter adjustment: Adjustment of the most important parameters for edge detection
- Real-time evaluation: Real-time analysis including visualization

The workflow is essentially as follows: A camera takes pictures of the tensile test specimen and periodically sends them to the defined endpoint of the system via POST request. First, a checkerboard pattern is clamped into the system for calibration. The user can then set the calibration parameters in the system's graphical user interface (GUI) to correct lens distortions in the area of the checkerboard pattern, thus facilitating subsequent edge detection. The operator then adjusts the parameters for edge detection in the GUI to ensure the optimal setting for the respective test. Once these parameters have been determined, the live visualization in the GUI can be used to track the detected sample geometry, its changes, and the changes in the associated control parameters (voltage).

User interface

For practical application, an intuitive graphical user interface was developed using Python and the PyQt6 framework [1]. The application is divided into three functional areas, which are accessible via separate tabs.

The Live View tab enables real-time monitoring of the measurement process with continuous display of the camera image and the detected edges. The parameter settings tab provides access to all relevant algorithm parameters such as Canny thresholds, RANSAC parameters, and curvature filter settings to dynamically adapt the methods used to new environments, lighting conditions, and samples. Finally, the calibration tab provides an interface for camera calibration for distortion correction to reduce inaccuracies in the fitting process.

Camera calibration

Correcting camera-induced distortions is important for precise measurements. The system uses proven OpenCV calibration functions to automatically determine and correct distortions in the camera optics [2]. A standardized checkerboard pattern with precisely known geometric properties is used as the calibration object.

The calibration process is performed by automatically detecting the corners of the checkerboard. From these point correspondences, the algorithm calculates the intrinsic camera parameters and the extrinsic parameters for each image. The determined distortion parameters are then used to correct all measurement images, eliminating systematic measurement errors caused by optical distortions.

Figure 1 shows how the distortion in the area of the tensile specimen is corrected using the GUI with the aid of a calibration sample.

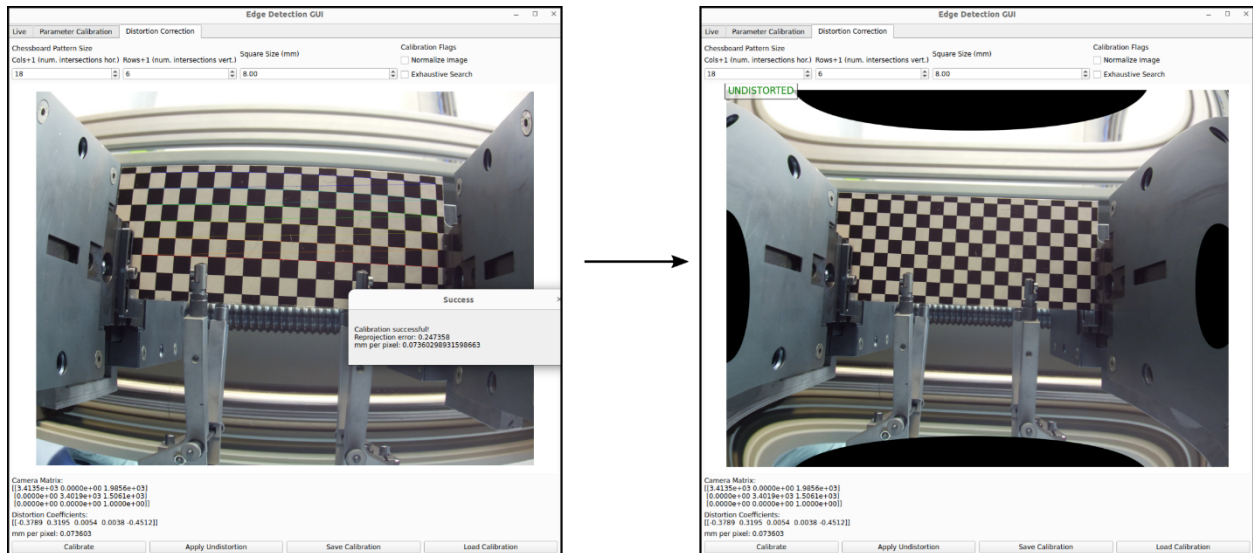


Figure 1. Calibration of the camera in the GUI using a checkerboard pattern

Edge detection and control parameters

Overview

The developed method consists of the following components: Canny edge detection with subsequent curvature filtering, robust line detection using the RANSAC algorithm, distance calculation to determine the sample width, and final mapping of the width change to process parameters for control applications. These methods are executed sequentially to extract distance measurements from an input image and convert them into controllable output signals.

Region of interest definition

To reduce computing time and focus on relevant image areas, a region of interest (ROI) is first defined. This rectangular region is defined by two corner points $P_1 (x_1, y_1)$ and $P_2 (x_2, y_2)$ and limits the working area of the algorithm to the area in which the sample edges to be detected are expected.

The ROI selection is based on a priori knowledge about the application and can be dynamically adjusted by the user in the GUI. This not only conserves computing resources, but also eliminates interference from other, irrelevant image areas. An example of an ROI selected in this way is shown in Figure 2.

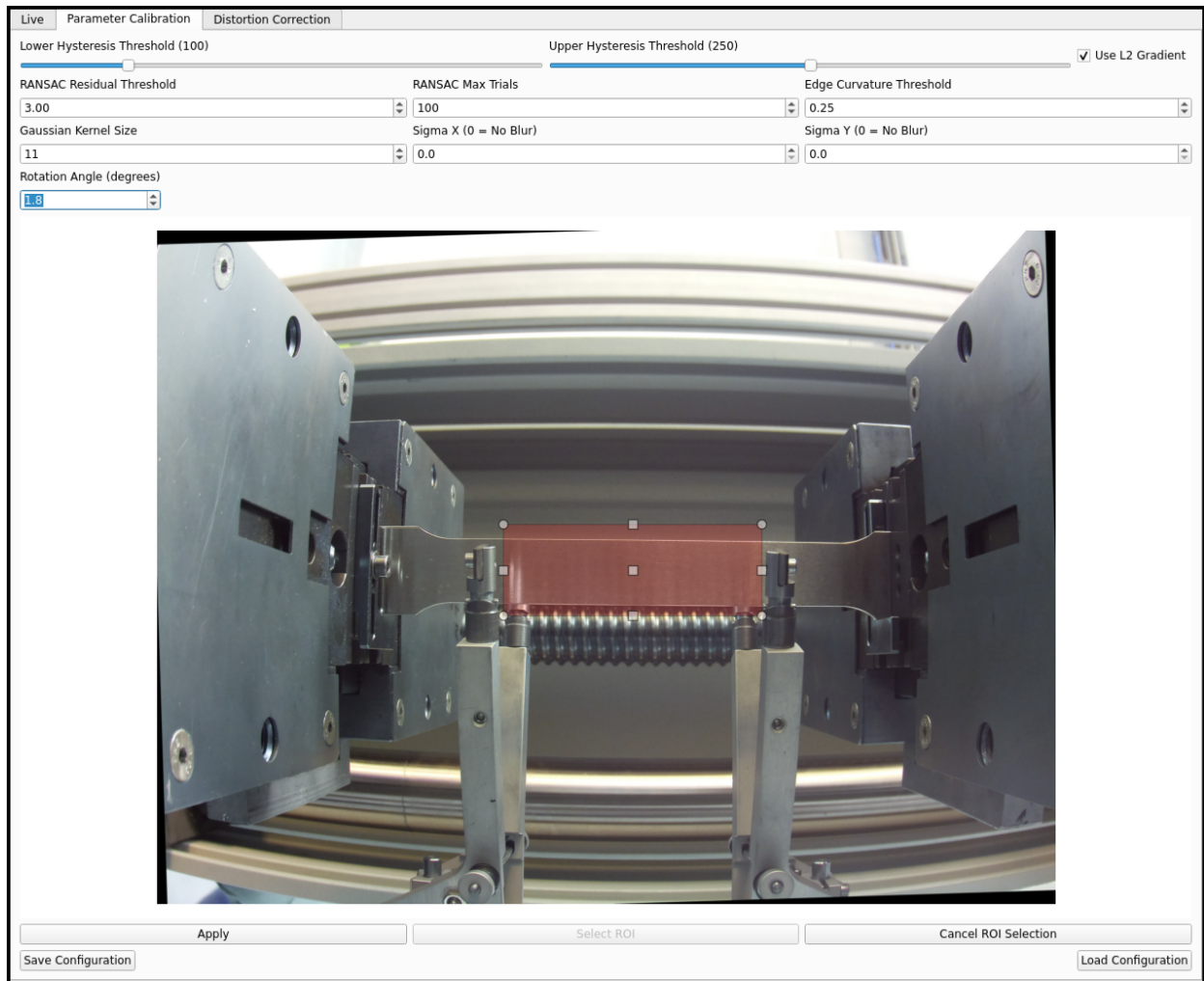


Figure 2. Definition of the region of interest (red rectangle) in the GUI. Only this image area is evaluated for edge detection.

Canny algorithm

The proven Canny edge detector, which is characterized by high accuracy, among other things, is used for edge detection [3]. The algorithm works with two thresholds: a lower and an upper threshold, which determine, among other things, how strong the brightness gradient at a point must be in order to be classified as an edge.

Pixels with gradients above the upper threshold value are immediately classified as edge pixels, while pixels between the two threshold values are only considered edges if they are connected to already identified edge pixels. This hysteresis thresholding results in closed edge lines and reduces fragmented edges.

Curvature-based filtering

In order to distinguish and filter out straight lines from curved structures in a naive first instance, curvature filtering is applied to the detected edges after Canny edge detection using polynomial fitting. For each contour, the x and y coordinates are extracted and a third-degree polynomial of the following form is fitted:

$$y = a_3x^3 + a_2x^2 + a_1x + a_0$$

The coefficients a_3 and a_2 characterize the curvature properties of the line. The maximum curvature value is defined as the maximum of the absolute values of these two coefficients:

$$\kappa = \max(|a_2|, |a_3|)$$

Contours with a curvature value below this defined maximum value, which can also be configured in the GUI, are classified as straight lines and retained for further processing. This filtering step is a first naive step to eliminate obviously curved structures and focuses the analysis on the desired straight edges.

RANSAC-based line detection

The edges of the tensile specimens are assumed to be two linear straight lines. The robust detection of these lines and the determination of the respective linear equations is performed using the RANSAC (Random Sample Consensus) algorithm, which is particularly resistant to outliers [4]. Since two horizontal (parallel) lines are expected in the application, the ROI is divided horizontally into an upper and lower half, and a straight line is fitted separately for each half using RANSAC, representing the upper and lower sample edges, respectively.

The RANSAC algorithm works iteratively: In each iteration, a random subset of at least two points is selected and a line is drawn through these points. The number of inlier points that lie within a defined residual threshold to the fitted line is then determined. After a maximum number of attempts, the model with the most inliers is selected as the best solution.

Each line is then described by the slope k and the y-intercept d . For the lower half of the image, the y-intercept is corrected by half the image height to obtain the correct position in the overall image.

Distance calculation

The distance between the two detected lines is measured at the horizontal center of the ROI ($x^{mid}=w/2$, where w represents the width of the ROI). The y-coordinates of both lines are calculated at this position: $y_1^{mid} = k_1 \cdot x^{mid} + d_1$

$$y_2^{mid} = k_2 \cdot x^{mid} + d_2$$

The distance between the lines is calculated as the absolute value of the difference between these y-coordinates:

$$d = |y_2^{mid} - y_1^{mid}|$$

Figure 3 shows the result of edge detection performed in this way, including the calculated distance (in pixels), with lines marked in red representing edges excluded by RANSAC.

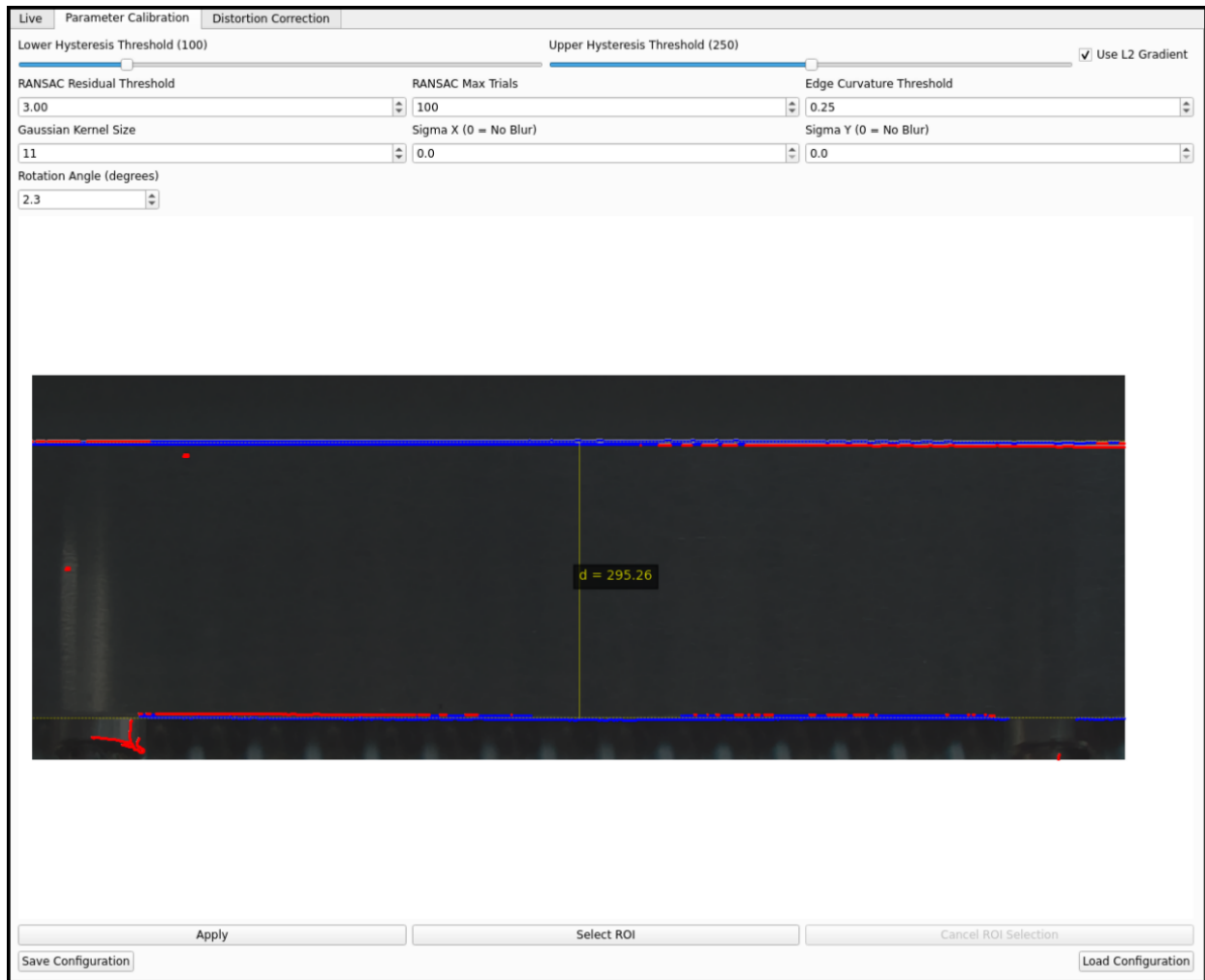


Figure 3. Result of edge detection. The blue lines represent the edges classified as sample edges.

Mapping for downstream control application

For integration into control systems, the measured distance d is converted into a voltage signal. For this purpose, a reference distance d' (before pulling the sample) and a corresponding reference voltage U' are defined.

The percentage deviation from the reference distance is then calculated as:

$$\Theta = \frac{(d - d')}{d'} \cdot 100$$

Using this deviation, a new voltage value can then be specified using a scaling factor α , which can also be configured in the GUI, and this value can then be transmitted to the control application:

$$U_{new} = U' + \Theta \cdot \alpha$$

Figure 4 shows the live view of the GUI that is displayed while the sample is being pulled. It shows both the detected sample edges and the calculated change in the edge distance from the reference distance before pulling. In addition, a factor for voltage mapping can be set directly in the GUI to convert the distance change linearly into a control variable to be output.

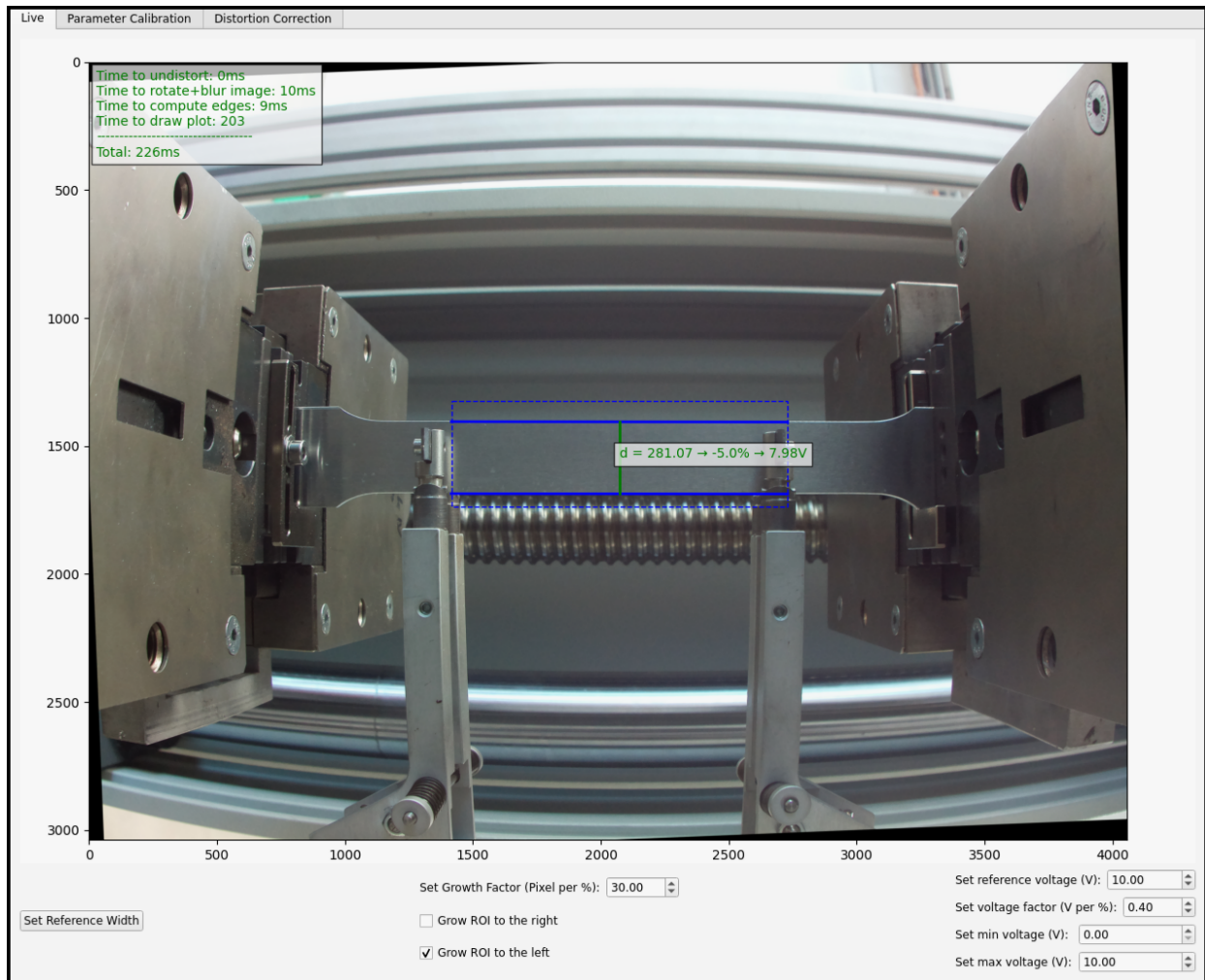


Figure 4. Shows the live view in the GUI during the drawing of a tensile sample, including detected edges in blue solid lines as well as line spacing in pixels and percentage change compared to the reference value before drawing and the resulting control variable (stress)

Results

The system was successfully validated at the LKR's semi-industrial plant. The plastic deformation of the sample was identified and quantified in real time by the optical module. Based on the recorded thinning values, the current control path was dynamically adjusted by feeding mapped control parameters into the high-current source control. This made it possible to specifically delay crack formation and improve formability.

The combination of robust, image-based geometry measurement and low system latency makes the solution particularly attractive for small and medium-sized companies with variable forming processes and limited sensor use.

The innovation of the solution lies in the implementation of real-time, image-based closed-loop control of the electroplastic effect (EPE) in metallic forming processes. Through optical detection of plastic thinning using Canny and RANSAC analysis, the current pulse control is dynamically adjusted so that the EPE-induced formability can be specifically influenced. The solution operates with low latency on edge hardware, is modularly scalable, and forms the basis for future AI-supported control systems in multi-axis deep drawing processes using digital image correlation (DIC).

In a planned follow-up project, the system is to be expanded for complex, multi-axis deep drawing processes. The use of digital image correlation (DIC) for multi-axis strain measurement is planned,

supplemented by a generic, adaptive control module on edge hardware. The goal is to develop a self-optimizing, adaptive control system for the intelligent operation of future forming systems.

Summary

The electroplastic effect (EPE) enables the reduction of forming forces and increases ductility in metallic materials. However, previous methods worked with rigid current profiles without feedback on the actual plastic deformation. This paper develops the basis for a real-time, image-based closed-loop control system that enables adaptive control of current pulses via optical detection of the sample geometry.

The system integrates camera calibration, Canny edge detection, RANSAC-based line detection, and real-time visualization in a GUI application. Continuous monitoring of sample thinning, which was developed in this paper, allows control parameters to be dynamically adjusted and fed into the high-current source. The combination of robust image analysis and low system latency enables precise control interventions during the forming process. Successful validation at the LKR's semi-industrial facility confirmed the practical suitability of the system, whereby deformation could be identified and quantified in real time.

The developed solution can also be transferred to other branches of industry: The principle of image-based closed-loop control can be applied to various forming processes, with small and medium-sized enterprises benefiting in particular from the cost-efficient edge hardware solution without complex sensor technology. Future AI-supported enhancements with digital image correlation open up potential for self-optimizing production systems in a wide range of manufacturing areas.

References

- [1] Computing, Riverbank, „PyQt6,“ [Online]. Available: <https://www.riverbankcomputing.com/software/pyqt/>. [Zugriff am 28 07 2025].
- [2] „OpenCV - Github,“ [Online]. Available: <https://github.com/opencv/opencv-python>. [Zugriff am 28 07 2025].
- [3] J. Canny, „A Computational Approach to Edge Detection,“ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bde. %1 von %2PAMI-8, Nr. 6, pp. 679-698, 1986.
- [4] R. C. B. Martin A. Fischler, „Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,“ *Communications of the ACM*, Bd. 24, Nr. 6, pp. 381-395, 1981.

Digital human-centered assistance systems

Christoph Heindl
Visual Computing
PROFACTOR GmbH
cheind@profactor.at

Gerhard Ebenhofer
Visual Computing
PROFACTOR GmbH
gebenh@profactor.at

Gernot Stübl
Visual Computing
PROFACTOR GmbH
gstueb@profactor.at

Harald Bauer
Visual Computing
PROFACTOR GmbH
hbauer@profactor.at

Kurzzusammenfassung

Der zunehmende Fachkräftemangel, steigende Variantenvielfalt und der Wunsch nach ergonomischen Arbeitsplätzen erfordern neue Ansätze für mensch-zentrierte Automatisierungslösungen. In diesem Beitrag wird eine modulare Toolbox vorgestellt, welche KI-basierte Wahrnehmung, projektive Interaktion und Robotik kombiniert, um adaptive Assistenzsysteme in variantenreichen Produktionsumgebungen zu realisieren. Zentrale Bestandteile sind die kamerabasierte Erkennung von Objekten und Personen sowie orts- und zeitgenaue Projektionen zur direkten Arbeitsplatzunterstützung. Für die Auftragserstellung stehen intuitive Modalitäten zur Verfügung: wie die Stifteingabe oder SketchGuide+, ein freihändiges Skizziersystem mit Wearable und visueller Rückmeldung. Die praktische Machbarkeit wurde in mehreren TBI-Projekten in Branchen wie Möbelbau, Sanitär- und Metalltechnik oder Solarproduktion demonstriert. Die Ergebnisse zeigen hohes Potenzial für Prozessverbesserung, Benutzerfreundlichkeit und Flexibilität in manuellen oder teilautomatisierten Abläufen.

Keywords: Digitale Assistenz, Visuelles Szenenverständnis, Projektion, Robotik

Abstract

The increasing shortage of skilled workers, growing variety of products and the desire for ergonomic workplaces require new approaches to human-centred automation solutions. In this work, we present a modular toolbox that combines AI-based perception, projective interaction and robotics to implement adaptive assistance systems in diverse production environments. Key components include camera-based recognition of objects and people, as well as location- and time-specific projections for direct workplace support. Intuitive modalities are available for order creation, such as pen input or SketchGuide+, a freehand sketching system with wearable technology and visual feedback. The practical feasibility has been demonstrated in several TBI projects in industries such as furniture manufacturing, sanitary and metal technology, and solar production. The results show significant potential for process improvement, user-friendliness, and flexibility in manual or semi-automated processes.

Keywords: Digital Assistance, Visual Scenes, Projection, Robotics

Introduction

The increasing complexity of industrial processes, the growing demand for product customization, and the ongoing shortage of skilled workers are presenting manufacturing companies with ever greater challenges. Human-centered assistance systems offer promising approaches to making work processes more flexible, ergonomic, and efficient without displacing humans from the center of the activity.

Existing assistance systems often focus on highly structured scenarios in which, for example, projection patterns are displayed according to rigid procedures and no adaptation to the working methods of workers takes place. They also often require complete automation, which is rarely found in small and medium-sized enterprises. In addition, the screens or touch panels are usually spatially and functionally separated from the real work context, which causes a high cognitive load for workers due to the media discontinuity. Although the first systems with augmented reality displays or sensor-based robotics already exist, their configuration and programming often requires specialized expertise. Integration into diverse, manual production environments often remains limited.

The approach presented here takes a different path: the modular toolbox developed, which combines AI-based perception, projective interaction technologies, and adaptive robotics, provides assistance where it is immediately needed, see Figure 1. In other words, directly on the workpiece, in the environment, and in the real workflow. Key elements include camera-based recognition of objects, people, and contextual information, as well as the location- and time-specific projection of recommended actions or input fields.

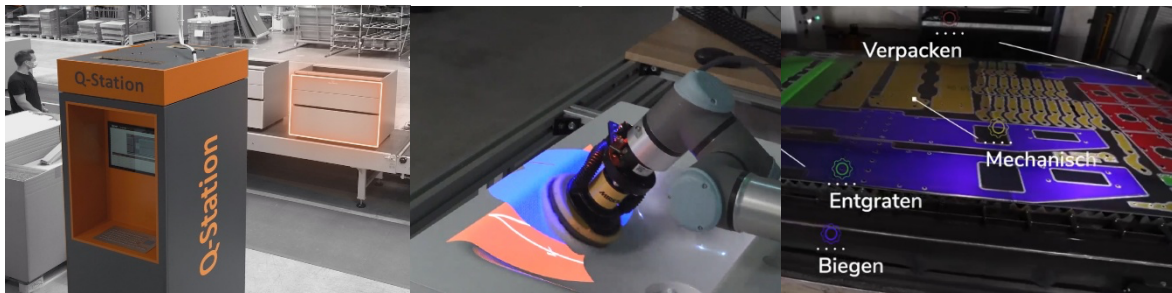


Figure 1. Diverse application scenarios for the presented toolbox for human-centered assistance systems: from projection-based assembly support and robot-assisted polishing and quality assurance to AI-supported object recognition and hands-free task definition with SketchGuide+.

Two equivalent, intuitive modalities are available for order entry: classic pen input with projective feedback and SketchGuide+ [1,2], a system for freehand sketching on real surfaces, enhanced with a portable wearable device for real-time parameter input and time-synchronized projections. Both methods enable low-threshold interaction, even for non-technical users, and allow rapid adaptation to changing requirements.

The technical and ergonomic feasibility of this meta-approach has been validated in several industry-specific short projects. In areas such as furniture construction, sanitary ware production, solar frame manufacturing, and logistics, potential for improved process quality, reduced setup times, and increased user acceptance has been demonstrated.

Approach focusing on key risk issues

The methodological approach in the successful TBI projects followed a structured, practical scheme aimed at investigating the greatest technological uncertainties at an early stage and under realistic conditions. The focus was not on the complete implementation of an overall system, but on the targeted validation of risky core issues with high relevance for later implementation. The approach included:

1. Early data collection in real operation:
Right from the start of the project, real parts data and process information were recorded directly on site in order to gain a robust understanding of the actual application context.
2. Focus on critical feasibility aspects:

Instead of broad requirements analyses, the respective problems were consistently reduced to a few key uncertainties, such as the quality of a robotic polishing process, the robustness of an AI model, or the intuitive usability of an interaction interface.

3. Targeted selection and adaptation of toolbox components:
Depending on the risk profile, suitable modules were selected from the toolbox (e.g., AI-supported image processing, projection-based assistance, interaction via pen or SketchGuide+) and adapted to the specific project, for example, by fine-tuning artificial neural networks or integrating hardware.
4. Lean system integration:
To limit the effort involved, the connection to operational information systems was kept as simple as possible, sufficient for realistic testing without triggering complex IT projects.
5. Demonstrator setup and evaluation:
The resulting prototypes were installed either in the laboratory or directly at the partner company and evaluated qualitatively and quantitatively under real conditions, with a focus on target variables such as process quality, interaction time, fault tolerance, and user acceptance.

This risk-oriented approach made it possible to address key implementation hurdles at an early stage, make reliable statements about technological feasibility, and at the same time create a sound basis for follow-up projects or concept development.

Results

Unlike conventional assistance systems, which often rely on isolated interaction methods, rigid automation logic, or purely screen-based feedback, the approach presented here takes a holistically integrated approach: The toolbox combines AI-based environmental perception, adaptive robotics, and projection-based, context-sensitive human-machine interaction into a flexible modular system for human-centered assistance systems.

The key innovative features are:

1. Direct interaction in a real-world work context:
Instead of a virtual interface, information is displayed where it is needed, i.e., directly on objects, surfaces, or workpieces.
2. Multimodal order entry via pen or SketchGuide+:
Users can define robot paths either by classic pen input or freehand on the real surface using SketchGuide+, including portable parameter input and immediate visual feedback.
3. AI-supported real-time perception for diverse objects and scenarios:
Object shapes, packages, or order layouts are automatically recognized, analyzed, and prepared for robotic actions without explicit programming.
4. Easy technology transfer via TBI
The toolbox has been systematically tested in various domains such as sanitary engineering, furniture construction, logistics, and solar production, proving both technological and ergonomic feasibility.

This modular combination results in a powerful assistance system that can also be used for single-item production, frequent variant changes, and for non-technical users, with high adaptability, minimal setup effort, and intuitive usability.

In particular, our technologies were evaluated in the following tasks:

- A semi-automated polishing process for sinks was implemented using robotics, 3D scanning, and intuitive order entry (pen or SketchGuide+) to improve ergonomics and quality in single-piece processing.
- For situational adhesive application in furniture construction, a robot-assisted system with image processing and user-friendly input (touch or SketchGuide+) was implemented, enabling individual orders to be completed in seconds.
- An AR-based unloading system projects visual instructions directly onto sheet metal parts, reducing paper processes and improving efficiency in sorting and removal.
- The welding process for individual solar collector frames was semi-automated through robot guidance and path input via pen or SketchGuide+, significantly reducing setup time for one-off productions.
- Digital projections replaced physical templates in boat building, significantly reducing application times and storage costs.
- An AI-based system for package classification and tracking demonstrated that visual recognition works reliably even under difficult lighting conditions and in real time.
- The manual production of mineral cast washbasins was successfully supported by a collaborative robot system with user-defined processing logic and validated in a real production environment.

Table 1 groups the technologies by component.

Table 1. Toolbox: Technologies in practical use..

Toolbox component	Exemplary applications
<i>Object and human recognition with camera and AI</i>	<i>Image processing and segmentation for individual parts (adhesive application, package acceptance, polishing)</i>
<i>Projection technology for AR interaction</i>	<i>Digital placement templates, visual unloading instructions, projected welding paths</i>
<i>Robotic assistance systems</i>	<i>Semi-automated polishing, welding, finishing, adhesive application</i>
<i>Multimodal interaction (touch/pen/hand guidance)</i>	<i>Operation via pen input, touch-based interfaces, gesture-based programming</i>
<i>Digital assistance & visual feedback systems</i>	<i>Projection, API interfaces, integration into real processes</i>

Figures 2 to 4 show selected images from specific TBIs.

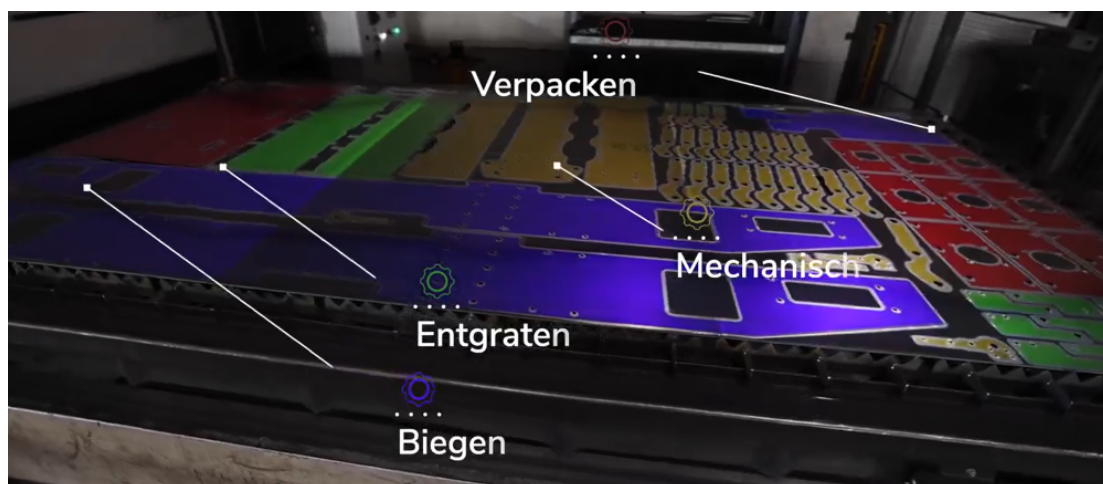


Figure 2. Digital unloading assistant in use: Projection of visual unloading instructions directly onto sheet metal parts to support manual sorting processes, enabled by a multi-projector system and lightweight integration into operational processes.

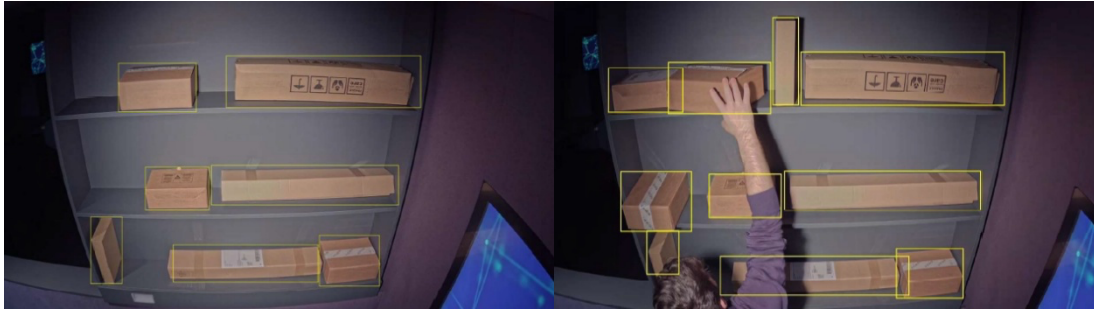


Figure 3. Digital assistant for parcel acceptance: AI-based object recognition reliably identifies and tracks parcels even in poor lighting conditions. A lean solution for open shelving systems without manual assignment.



Figure 4. Digital placement templates in boat building: Projection replaces physical templates and enables precise, time-efficient placement of rod decks directly on the work surface.

Summary

A toolbox for human-centered assistance systems was presented, combining AI-based perception, robotics, and projection-based interaction. It was used in several TBI projects in a practical setting to validate risky core issues at an early stage. Through real data collection, targeted adaptation of components, and lean prototypes, technological feasibility was demonstrated in various domains. In particular, the combination of context-sensitive projection and intuitive input, via pen or SketchGuide+, proved to be the key to flexible support for manual processes and to a well-founded assessment of implementation potential.

Unlike other solutions, the technology presented is not limited to use in large companies, but also brings small and medium-sized enterprises direct time savings and increased quality assurance by supporting manual activities. The flexibility of the toolbox makes it possible to adapt to either changed manufacturing requirements or the working methods of the workers. This is a clear advantage over other systems. The technology is designed to be industry-independent and is aimed at the still predominantly manual workplaces in Austrian production environments.

References

- [1] Heindl, Christoph, et al. "SketchGuide: A Baseline Vision-Based Model for Rapid Robot Programming via Freehand Sketching on Any Surface." European Robotics Forum. Cham: Springer Nature Switzerland, 2025.
- [2] Heindl, Christoph, et al. "SketchGuide+: Enabling Parametric Freehand Robot Programming via Wearable-Assisted Sketching." ACM International Joint Conference on Pervasive and Ubiquitous Computing, Espoo, Finland, 2025.

Intelligent assistance system for flexible human-robot collaboration in manufacturing

Helmut Nöhmayer
PROFACTOR GmbH
helmut.noehmayer@profactor.at

Christian Wögerer
PROFACTOR GmbH
christian.woegerer@profactor.at

Oswald Bratu
PROFACTOR GmbH
oswald.bratu@profactor.at

Andreas Pichler
PROFACTOR GmbH
andreas.pichler@profactor.at

Kurzzusammenfassung

Der steigende Bedarf an individualisierten Produkten, ergonomischen Arbeitsplätzen sowie der zunehmende Fachkräftemangel erfordern neue humanzentrierte Automatisierungsansätze. Dies gilt besonders für Anwendungen mit hoher Variantenvielfalt und geringer Stückzahl. In diesem Beitrag wird ein intelligentes Assistenzsystem vorgestellt, das KI-unterstützte Wahrnehmung, adaptive Robotik und multimodale Interaktionstechnologien kombiniert. Ziel ist es dabei, die Mensch-Roboter-Kollaboration (MRK) auch in manuell geprägten Fertigungsprozessen praktikabel und wirtschaftlich zu gestalten. Das System ermöglicht die intuitive Auftragserstellung mittels Stifteingabe, Gestik oder Sprache direkt am Werkstück. Durch die situative, automatische Generierung und Optimierung von Roboterbahnen können selbst Einzelstücke effizient bearbeitet werden – ohne klassischen Programmieraufwand. Die technische Machbarkeit wurde im Rahmen mehrerer industriebezogener Anwendungsfälle validiert. Die Ergebnisse zeigen signifikante Potenziale für die Reduktion von Rüstzeiten und bei Steigerung der Produktqualität, Ergonomie und Arbeitsplatzattraktivität

Keywords: Mensch-Roboter-Kollaboration, Assistenzsystem, KI-basierte Prozessplanung, Stifteingabe, flexible Automation

Abstract

The increasing demand for customized products, ergonomic workplaces, and the growing shortage of skilled labor call for new human-centered automation approaches. This is particularly relevant for applications characterized by high product variability and low production volumes. This paper presents an intelligent assistance system that integrates AI-based perception, adaptive robotics, and multimodal interaction technologies. The objective is to make human-robot collaboration (HRC) both practical and economically viable, even in manufacturing processes that are predominantly manual.

The system enables intuitive task specification directly at the workpiece through pen input, gestures, or speech. By automatically generating and optimizing robot trajectories in situ, even single-piece production can be carried out efficiently—without the need for traditional programming. The technical feasibility has been validated through several industry-related application scenarios. The results demonstrate significant potential for reducing setup times while improving product quality, ergonomics, and workplace attractiveness.

Keywords: Human-Robot Collaboration, Assistance System, AI-Based Process Planning, Pen Input, Flexible Automation

Introduction

In industrial manufacturing, there is growing pressure to respond flexibly to individual customer requirements while simultaneously ensuring and improving quality, efficiency, and ergonomics. Processes such as polishing, gluing, or welding are often characterized by monotonous, manual work,

especially in single-item or small-batch production. These activities are not only time-consuming and labor-intensive, but also pose health risks and are increasingly difficult to staff.

Traditional automation systems offer only limited solutions and the necessary operational flexibility: they sometimes follow rigid programming logic and thus require long setup times in relation to process times. In addition, they require special expertise. In practice, this means that many manual processes have not yet been automated—despite obvious potential for efficiency gains—because the programming times sometimes exceed the time required for manual execution by a considerable margin.

This article presents an intelligent, modular assistance system that enables a new level of human-robot collaboration through AI-based perception, sensor-supported clamping detection, and intuitive human-machine interaction. The focus is not on complete automation, but rather on the targeted partial automation of monotonous process steps while retaining human control and experience in terms of execution quality and results.



Figure 1. Order generation for spot processing of washbasins

Technological objective

The system aims to radically reduce order generation times during variant creation. Through the use of a multimodal user interface and situational visualization of the order and its progress (pen, gestures, voice, touch, projection), processing orders can be generated very quickly directly on the actual workpiece in a situation-specific manner. This is made possible by:

- Position and geometry recognition using 2D +3D camera systems with composite images of large-area/large-volume workspaces
- AI-supported recording of individual part clamping/part provisioning
- Intuitive and rapid input of application points and machining regions via hand gestures and tracked pointing instruments for individual spot machining
- Automated path planning and process generation for these situational machining jobs
- Projection-supported visualization and feedback for validating the machining paths directly on the product

The combination of these methods and components enables orders to be created within a few seconds per processing point (spot) – even for unique applications and batch size 1 products, see Figure 1. The actual order generation is fully automated in the background. This makes the use of robotics practicable for the first time even in environments with a high degree of variety and low automation expertise. Economically, this approach supports the reduction of setup costs/machine downtime and thus serves to lower unit labor costs with regard to competitiveness, see Figure 2.

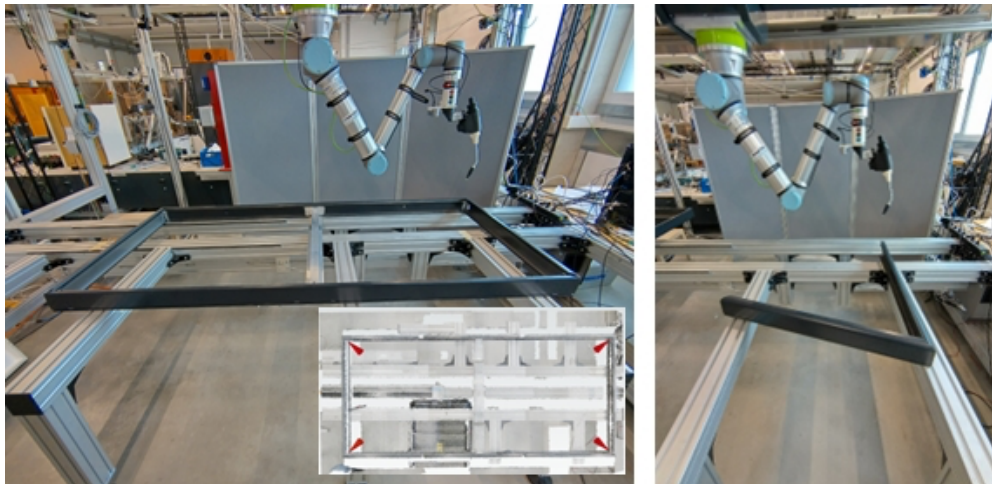


Figure 2. 3D capture of collector frames

Approach: Validation of risky core aspects

The development and validation of the system followed specifically identified industrial challenges in several TBI projects. The methodological approach was geared toward analyzing and evaluating the technical risks and deriving and assessing the conceptual limitations, focusing on the following key questions:

- Polishing processes:
 - Is it possible to achieve consistent high-gloss quality with a robot when working with a wide variety of sink shapes?
 - Is it possible to automatically capture the loading situation in 3D in such a way that processing regions for programming/order generation can be identified even for different sink shapes?
- Welding processes:
 - Can the geometry of individual solar frames be recognized automatically and used to derive a robust individual welding job in real time?
 - What sequence of position detection and position corrections is necessary for ad hoc welding process input for individual components, such as the angled mountings of solar frames for installation on roofs?
- Order input:
 - How can intuitive order generation for a unique product/process be implemented without programming knowledge in such a way that it can be carried out in a minimum amount of time and is particularly easy to learn?
- To answer these questions, realistic functional models were set up, equipped with collaborative robots, instrumented tools, 3D camera systems, and projection-based feedback components, among other things (see Figure 3).
- These systems were evaluated both in the laboratory and in real production environments.

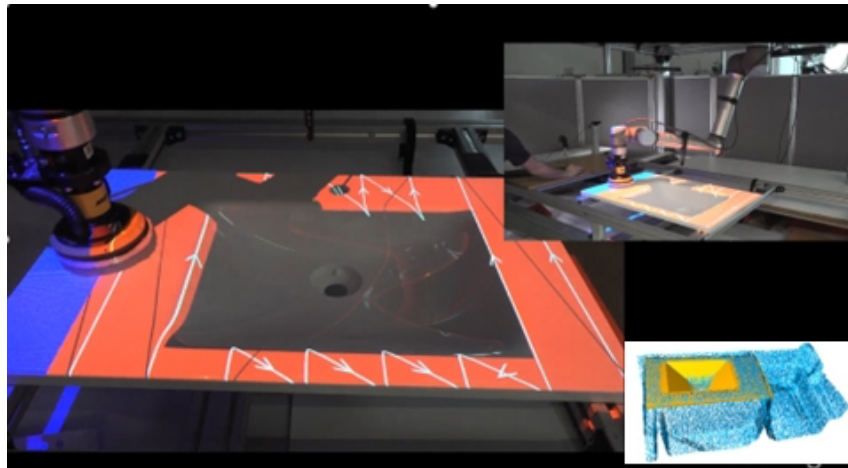


Figure 3. 3D capture of washbasins

Results

The validation results show that the developed system is capable of meeting key requirements for intuitive use of robot systems for unique processes:

- Setup time savings: Depending on the application and process, order entry times range from 10 to 30 seconds per process point.
- Quality: The execution of the machining process by a robot system delivers more consistent results even with varying geometries and clamping arrangements.
- User-friendliness: Users rated the intuitive interaction via touch, pen, gestures, and projection as easy to learn and apply.
- Scalability: The solution can be transferred to different manufacturing processes (polishing, grinding, welding, gluing, oiling, and even painting) without fundamental adjustments.

The technical feasibility has thus been successfully demonstrated under real conditions. Particularly noteworthy is the combination of real-time geometric recognition and intuitive order creation, which enables cost-effective automation even for single pieces.

Innovation and system architecture

The degree of innovation of the assistance system lies in particular in the consistent integration of the following sub-technologies:

- • 2D/3D storage recording of loading situations from composite images of large workspaces [1].
- • Instrumented tools and sensors for entering position, location, and processing intention.
- • Multimodal interaction: Combination of touchscreen, pen input [2], voice control, gesture recognition, and guidance/feedback via projection.
- • AI-supported process planning: Automatic conversion of user interactions into robot-assisted, situation-dependent action paths [3][4].
- • Adaptive parameterization: Automatic adaptation to material, geometry, and process status.

The interaction of these components creates a system that replaces classic programming logic with situational action – and thus initiates a paradigm shift in industrial robotics. This solution approach addresses in particular the needs of SMEs in terms of tapping automation potential for continued cost-effective manufacturing in Europe.

Benefits and implications for manufacturing

The developed solution offers a wide range of benefits for manufacturing companies, especially in the context of skilled labor shortages, batch size 1 manufacturing, and digital transformation:

- Wide applicability in typical manual machining processes in manufacturing in the areas of wood, metal, and plastics processing, such as grinding, brushing, polishing, oiling, painting, welding, or gluing.
- Interactive, rapid order generation for robot systems for product variants and process variations that were previously determined exclusively by humans and their expertise—e.g., rapid entry of a grinding order for intermediate grinding after assessment by experts, thereby reducing machine downtime and enabling more effective use of plant investment.
- Reduction of process costs by eliminating manual processing time and setup effort
- Increased flexibility with consistent quality
- Relieving the burden on employees through ergonomic semi-automation, thereby improving the workplace environment
- Strengthening the “human-machine” team to increase competitiveness and profitability
- Opening up new markets through the automation of even highly customized products
- Strengthening resilience through greater process stability in the face of varying requirements

The intuitive operation and low barriers to entry provide small and medium-sized enterprises (SMEs) in particular with realistic access to industrial robotics – even without their own automation expertise. Thanks to its technological combination, the developed and validated generalized solution approach can be used in many areas of manufacturing. Depending on requirements and processes, the range of functions can be designed in a modular fashion – from simple user guidance and projection to interactive programming of situational process steps on the component itself. This also allows users to deploy these sub-technologies in various forms at different production sites.

Summary

The assistance system presented marks a significant advance in human-robot collaboration for industrial manufacturing. The intelligent combination of sensor technology, AI-supported path planning, and multimodal interaction enables the economical use of robotics even for one-off productions and highly varied processes. Validation in real-world applications shows that intuitive, semi-automated programming directly on the workpiece is not only technically feasible, but also a decisive lever for the competitiveness of future production systems.

References

- [1] Mitteramskogler, J. J., Widmoser, F., Ikeda, M., Pichler, A.; Homography-based industrial image stitching Insight Session, (ERF 2024)
- [2] Ikeda, M., Widmoser, F., Stübl, G., Zafari, S., Sackl, A., & Pichler, A. (2023, October). An Interactive Workplace for Improving Human Robot Collaboration: Sketch Workpiece Interface for Fast Teaching (SWIFT). In Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (pp. 191-194).
- [3] Anish Pratheepkumar, Markus Ikeda, Andreas Pichler, Markus Vincze, Towards Robotic 3D Surface Processing with Global Local Neural Region Descriptor Fields, (ERF2024)
- [4] Möhl, P., Pratheepkumar, A., Ikeda, M., Pichler, A.; Morphing Based Transfer of Demonstrated Surface Finishing (ISM24)

Open source cybersecurity anomaly detection

Alexander Szönyi
SBA-Research gGmbH
aszoenyi@sba-research.org

Andreas Ekelhart
SBA-Research gGmbH
aekelhart@sba-research.org

Markus Klemen
SBA-Research gGmbH
mklemen@sba-research.org

Kurzzusammenfassung

Die digitale Souveränität gewinnt zunehmend an Bedeutung. In einem Use Case mit einem Unternehmen mit kritischer digitaler Infrastruktur wurde untersucht, inwieweit Open-Source-Lösungen bestehende Technologien im Bereich Security Operations ergänzen oder ersetzen können. Der Fokus lag dabei auf der frühzeitigen Identifizierung potenzieller Angriffe und Betriebsstörungen. Zu diesem Zweck wurden Open-Source-Werkzeuge wie Wazuh und OpenSearch in eine produktionsnahe Umgebung integriert und unter realen Rahmenbedingungen evaluiert. Neben den technischen Integrationsaspekten wurden auch regulatorische Anforderungen und relevante Standards berücksichtigt. Ein besonderes Augenmerk galt der Verarbeitung großer Datenmengen. Die Ergebnisse zeigen sowohl Potenziale zur Erweiterung bestehender Sicherheitsarchitekturen durch Open Source als auch Herausforderungen bei Skalierung, Integration und Betriebsstabilität. Daraus ergeben sich praxisnahe Empfehlungen für den Einsatz von Open-Source-basierten Lösungen im industriellen Umfeld sowie in anderen Organisationstypen.

Keywords: Security Operations, Anomaly Detection, Open-Source Security, Digital Sovereignty

Abstract

Digital sovereignty is becoming increasingly important. In a use case involving a company with critical digital infrastructure, we examined the extent to which open-source solutions can complement or replace existing technologies in the area of security operations. The focus was on the early identification of potential attacks and operational disruptions. To this end, open-source tools such as Wazuh and OpenSearch were integrated into a production-like environment and evaluated under real-world conditions. In addition to the technical integration aspects, regulatory requirements and relevant standards were also taken into account. Particular attention was paid to the processing of large amounts of data. The results show both the potential for expanding existing security architectures through open source and the challenges of scaling, integration and operational stability. This has resulted in practical recommendations for the use of open source-based solutions in industrial environments and other types of organisations.

Keywords: Security Operations, Anomaly Detection, Open-Source Security, Digital Sovereignty

Introduction

Securing complex IT infrastructures is becoming increasingly important in digital value chains. Companies with critical digital infrastructure are facing growing demands to monitor and analyze their data streams in order to detect and ward off potential attacks and operational disruptions at an early stage. In addition to technological challenges, regulatory requirements such as the EU directives NIS1 and NIS2, as well as international standards such as ISO/IEC 27001, must also be taken into account to ensure an appropriate level of security.

Against this backdrop, a use case was conducted to examine the extent to which open source solutions can complement or even replace existing technologies in the field of security operations. The focus was

on anomaly detection, a central component of modern security architectures that enables deviations from normal operating conditions to be identified, thereby detecting potential attacks at an early stage. To answer this question, open source components such as Wazuh and OpenSearch were integrated into a production-like environment and evaluated under real-world conditions.

The use case under consideration is characterized by a very high volume of data, which must be continuously monitored from a cybersecurity perspective. The data to be processed includes both standardized network and system logs as well as application-specific data that is central to the company's core business. This includes protocols from firewalls, switches, and routers (via the Syslog protocol), network traffic data, provider-specific connection data, portal information, and endpoint telemetry data. Depending on the network architecture, this results in high data rates, which are typically measured in events per second (EPS). In this project, the processing of input data streams at a rate of more than 100,000 to 200,000 EPS was therefore investigated, which corresponds to a data volume of approximately 1 TB per day (see Table 1). Taking into account a retention period of 365 days, this results in storage and processing requirements in the range of several hundred terabytes.

Tabel 1. Estimated data volume per day.

Source	GB/day
Netflow data	100
Syslog data	150
Provider data	400
Endpoint data	100
Total	750

Powerful open source components suitable for setting up and operating modern security operations functionalities were used to implement the project. The focus was particularly on OpenSearch solutions for highly scalable storage and analysis of log data, and Wazuh for the collection, normalization, and correlation of security-related events. The aim of the project was to evaluate the performance of these technologies in a production-like environment, identify integration hurdles, and derive recommendations for action for the introduction of open-source-based anomaly detection platforms.

Problem-solving approach

A methodological approach based on a modular architecture and the use of virtualized open source components was chosen for the evaluation. The goal was to collect and preprocess different data sources under real conditions and make them available for security analyses. The focus was particularly on the scalability of data collection, load distribution across multiple instances, and reliable forwarding to analysis clusters. The solution developed is based on a modular architecture that integrates various open source components and enables flexible scaling (see Figure 1).

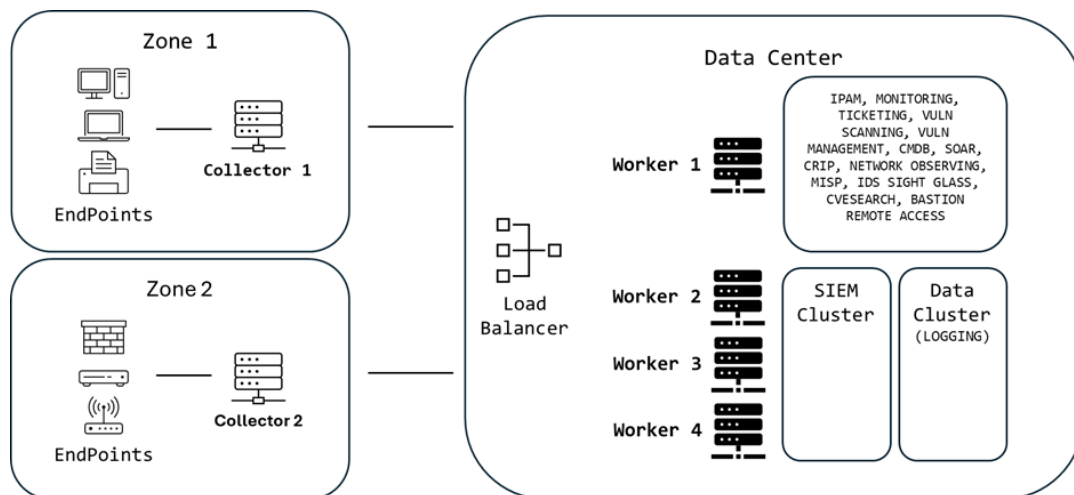


Figure 1. Example illustration of the open source-based SOC solution.

To implement the SOC functionalities, several open source modules were integrated, covering various tasks such as asset management, vulnerability analysis, and network monitoring (see Figure 2).

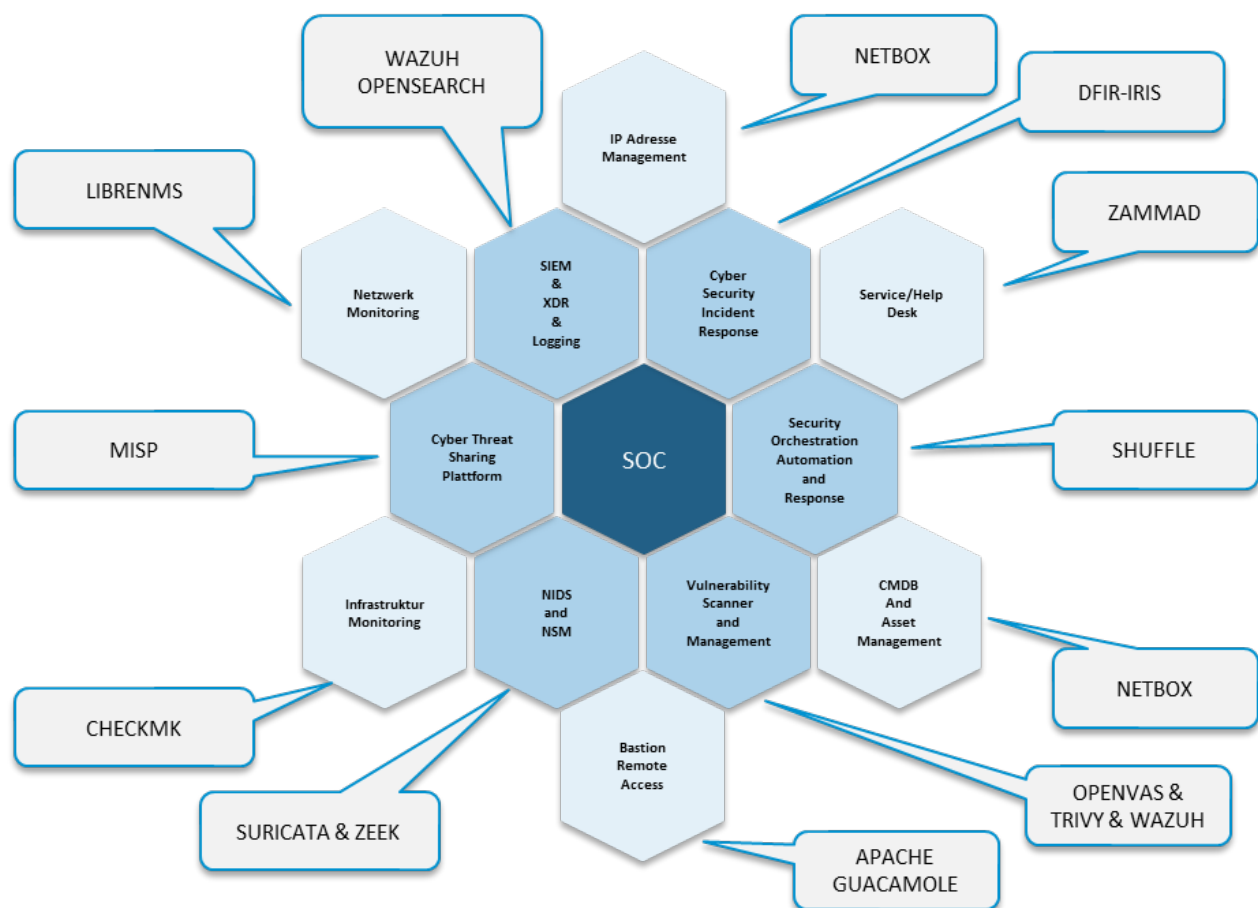


Figure 2. Illustration of the modules used in the open source-based SOC solution.

The following section describes the approach taken to process the various data sources.

Syslog data

A Wazuh agent was selected as the input collector for recording syslog data, as Wazuh already provides an integrated processing chain for these data sources. The agent receives the data streams and forwards

them to a Wazuh server cluster, which analyzes the events and prepares them for downstream security processes (see Figure 3).

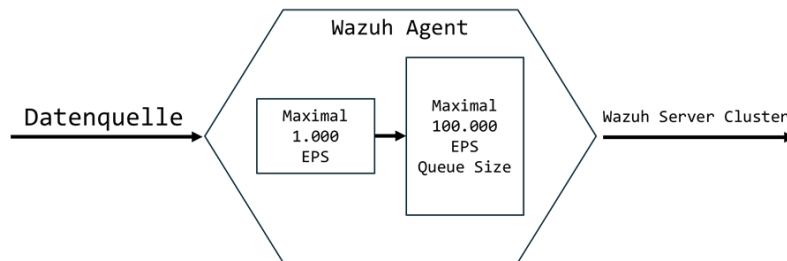


Figure 3. The illustration shows the data flow and limitations.

However, a single Wazuh agent is subject to performance limits in terms of the number of events per second (EPS) it can process and the maximum temporary storage of events (queue size). In order to be able to process a data rate of more than 100,000 EPS, several Wazuh agents were therefore operated in parallel within a virtualized container environment. An open-source-based load balancer was used for load distribution, which distributes the data stream to the individual agents via dedicated network ports (see Figure 4).

In test operation, up to 20 Wazuh agents running in parallel could be operated on a single host system without any data loss.

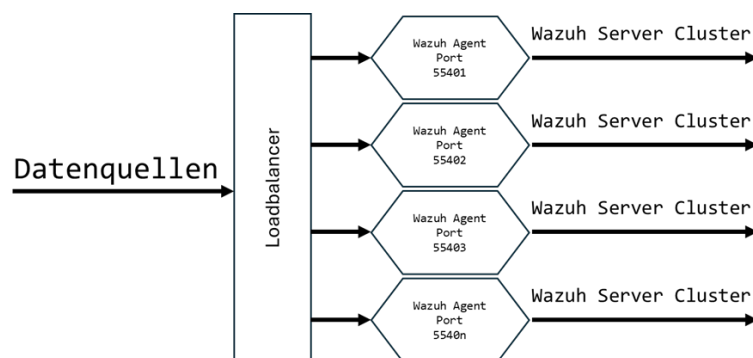


Figure 4. The diagram shows the data flow and the distribution of traffic across multiple agents..

Netflow data

An analog approach was chosen for processing network traffic data, in particular Netflow (IPv4/IPv6): virtualized Logstash instances served as receivers of the Netflow data streams and transferred the processed data to the OpenSearch cluster.

Provider data

Provider-specific connection and operating data was also collected and normalized using virtualized Logstash instances and integrated into the central OpenSearch data storage for further analysis.

Results

The evaluation showed that open source solutions are capable of processing very large amounts of data reliably and without data loss. A key contribution of the project was to combine various open source components, including Wazuh, OpenSearch, and virtualized Logstash instances, in a modular framework and adapt them to the specific requirements of the use case. The flexible architecture allowed different data sources to be integrated, processed, and evaluated in a scalable test environment under realistic conditions. It was demonstrated that both standardized and business-relevant individual data can be integrated with comparatively little integration effort.

A key feature of the solution stack developed is the complete on-premises processing of all security-related information, meaning that no data leaves the company's own data center. This ensures the digital sovereignty of the information—a crucial factor, especially for operators of critical infrastructures. The results achieved are in line with the goals of the current Austrian government program, which emphasizes strengthening innovation and digitalization and calls for the avoidance of unilateral technological dependencies. Building on the insights gained, the developed framework offers potential for further fields of application, such as the integration of additional data sources or the development of automated response mechanisms. In addition, the modular architecture opens up opportunities for cooperation with other organizations to establish common security standards and open reference architectures.

Summary

The project has demonstrated that integrating multiple open-source components can create a powerful anomaly detection framework that works reliably even at very high data rates and flexibly integrates individual, business-relevant data sources. A key advantage is the complete on-premises processing of all security-relevant information, which ensures digital sovereignty. The approach has other strengths as well: it is scalable for organizations of all sizes and industries, supports compliance with regulatory requirements (e.g., NIS1/2, DORA) and international standards (e.g., ISO/IEC 27001, ISO 62443), and enables rapid implementation thanks to its open architecture. The use of open-source solutions also strengthens the domestic economy and benefits from a global community that continuously drives further developments. This makes the results achieved highly relevant for a wide range of organization types and industries.

Several research priorities are emerging for the future. These include the normalization of ISP log data to create a standardized and consistent database, as well as the expansion of anomaly detection to include machine learning methods and specific endpoint profiling.

References

- [1] Wazuh, *Agent Client Buffer Einstellungen*. [Online]. Available: <https://documentation.wazuh.com/current/user-manual/reference/ossec-conf/client-buffer.html>. [Accessed: 30-Jul-2025].
- [2] OpenSearch, *Logstash Plugin*. [Online]. Available: <https://docs.opensearch.org/docs/latest/tools/logstash/index/>. [Accessed: 30-Jul-2025].
- [3] Heise Online, *Strafgerichtshof: Microsofts E-Mail-Sperre als Weckruf für digitale Souveränität*. [Online]. Available: <https://www.heise.de/news/Strafgerichtshof-Microsofts-E-Mail-Sperre-als-Weckruf-fuer-digitale-Souveraenitaet-10387368.html>. [Accessed: 30-Jul-2025].
- [4] United States Department of Justice, *Division V—Cloud Act*. [Online]. Available: <https://www.justice.gov/criminal/media/999391/dl?inline>. [Accessed: 30-Jul-2025].
- [5] Golem.de, *Microsoft kann US-Zugriff auf EU-Cloud nicht verhindern*. [Online]. Available: <https://www.golem.de/news/anton-carniaux-microsoft-kann-us-zugriff-auf-eu-cloud-nicht-verhindern-2507-198283.html>. [Accessed: 30-Jul-2025].

RAG is Not All You Need!

Raphael Pisoni
Software Competence Center Hagenberg
Raphael.pisoni@scch.at

Mario Pichler
Software Competence Center Hagenberg
Mario.pichler@scch.at

Kurzzusammenfassung

Retrieval Augmented Generation (RAG) hat sich zu einem Standardparadigma entwickelt, das es Large Language Models (LLMs) ermöglicht, Fragen unter Verwendung privater oder Echtzeit-Informationen zu beantworten. Durch das Abrufen relevanter Dokumente und deren Bereitstellung als Kontext für den Generator zielen RAG-Systeme darauf ab, faktenbasierte und spezifischere Antworten zu erzeugen. In diesem Beitrag wird jedoch argumentiert, dass der Standard-RAG-Ansatz für komplexe Informationsbedürfnisse unzureichend ist. Es werden zwei grundlegende Schwächen identifiziert: Erstens erfasst der Retrieval-Schritt oft nicht den vollständigen Kontext, der für eine genaue Beantwortung einer Frage erforderlich ist, was dazu führt, dass unvollständige oder irreführende Informationen an den Generator weitergegeben werden. Zweitens kann selbst bei erfolgreichem Retrieval das schiere Volumen des relevanten Textes die Kapazität des *Context Windows* des Generators überschreiten, was eine Kürzung der Informationen erzwingt und das Modell daran hindert, eine umfassende Antwort zu synthetisieren. Um diese Mängel zu beheben, wird eine neuartige Methode vorgeschlagen: Retrieval Augmented Context Engineering (RACE). Diese Methode führt zunächst ein breites Retrieval potenziell relevanter Textausschnitte durch. Anschließend fungiert ein generatives Modell als intelligenter Filter, der entweder irrelevante Dokumente durch die Erzeugung eines einzigen „None“-Tokens schnell verwirft oder prägnante, auf die Frage ausgerichtete Zusammenfassungen der relevanten Dokumente erstellt. Diese kompakten Zusammenfassungen, die alle für die Frage relevanten Informationen enthalten, können dann gemeinsam in das *Context Window* des finalen Generators eingespeist werden. Dadurch wird sichergestellt, dass der Generator Zugriff auf die gesamte Bandbreite der notwendigen Informationen hat, ohne dass es zu einem Überlaufen des Kontext Windows kommt, was qualitativ hochwertigere und vollständigere Antworten erzeugt.

Keywords: Large Language Models, Retrieval Augmented Generation, Industrielle Anwendungen

Abstract

Retrieval Augmented Generation (RAG) has become a standard paradigm for enabling Large Language Models (LLMs) to answer questions using private or real-time information. By retrieving relevant documents and providing them as context to the generator, RAG systems aim to produce more factual and specific answers. However, this paper argues that the standard RAG approach is insufficient for complex information needs. We identify two fundamental weaknesses: first, the retrieval step often fails to capture the complete context required to answer a question accurately, leading to incomplete or misleading information being passed to the generator. Second, even when retrieval is successful, the sheer volume of relevant text can exceed the generator's context window capacity, forcing a truncation of information and preventing the model from synthesizing a comprehensive answer. To address these shortcomings, we propose a novel method, *Retrieval Augmented Context Engineering (RACE)*. This method first performs a broad retrieval of potentially relevant text snippets. Then, a generative model acts as an intelligent filter, either rapidly discarding irrelevant documents by generating a single "None" token or producing concise, question-focused summaries of the relevant ones. These compact summaries, which retain all question-relevant information, can then be collectively fed into the final generator's context window. This ensures that the generator has access to the full breadth of necessary information without context overflow, which is designed to produce higher quality and more complete answers.

Keywords: Large Language Models, Retrieval Augmented Generation, Industrial Applications

Introduction

The advent of Large Language Models (LLMs) has revolutionized natural language processing, but their inherent knowledge is static, limited to the data they were trained on. Retrieval Augmented Generation (RAG) [1] was introduced as an elegant solution to this problem, allowing LLMs to access and utilize external knowledge sources at inference time. The typical RAG pipeline is straightforward: given a user query, a retriever (often based on vector similarity search) [2] fetches a set of documents deemed relevant from a corpus. These documents are then concatenated with the original query and fed as context to a generator LLM, which produces the final answer.

This approach has proven effective for many tasks, particularly for fact-based questions where the answer is explicitly contained within a few retrieved passages. However, as we move towards more complex, multi-faceted questions, the limitations of this simple pipeline become increasingly apparent. The quality of the final generation is critically dependent on the quality of the retrieval, a process that is far from perfect. Furthermore, the physical constraint of the LLM's context window imposes a hard limit on the amount of information the model can "see" at once.

This paper contends that for complex queries, RAG is not enough. We will first delve into the dual problems of incomplete context retrieval and context window overflow. In its place, we propose Retrieval Augmented Context Engineering (RACE), an architecture that enhances the RAG framework by introducing a generative filtering and summarization step. This intermediate process ensures that the final generation stage receives a context that is both complete and concise, containing all the necessary information to formulate a high-quality answer without being constrained by context length.

The Shortcomings of Standard RAG

The efficacy of a standard RAG system hinges on two core assumptions: (1) the retriever can successfully find all the relevant information, and (2) the generator can fit all this relevant information into its context window. In practice, both of these assumptions frequently break down.

The Incompleteness of Retrieval

Modern retrieval systems, while powerful, are imperfect. Vector-based search, the most common method, excels at finding semantic similarity but can struggle with nuanced queries. A query might require synthesizing information from multiple, disparate sources that do not explicitly mention each other or the exact keywords in the query. For instance, a question like, "What was the market's reaction to Company X's sustainability report, considering the new regulations announced last quarter?" requires retrieving at least two distinct types of documents: financial news about Company X and documents detailing the new regulations.

A standard RAG system might retrieve one or the other, but rarely both, especially if the documents do not directly cross-reference each other. The retriever, optimized for finding a few "top-k" relevant chunks, often misses pieces of the puzzle. This leads to a generator that is working with an incomplete picture, resulting in answers that are partially correct but lack the depth and synthesis the user is seeking [3, 4]. The model might correctly describe the market reaction but fail to connect it to the new regulations, thereby failing to answer the user's full question.

The Challenge of Context Window Size

Even if the retriever could perfectly identify every single relevant document, a second, more practical problem arises: the limited context window of the generator LLM. While context windows have been growing, they remain a finite and valuable resource. Complex questions can easily have dozens of relevant source documents or passages. Trying to concatenate all of this raw text into the prompt will quickly exceed the context limit of even the most advanced models.

The standard solution is to simply take the top-k retrieved documents that fit within the window. This truncation is a crude and often destructive form of compression. The (k+1)th document, which might contain a crucial piece of contradictory or clarifying information, is simply discarded. The model is forced to generate an answer based on an arbitrarily truncated subset of the relevant information. This can lead to answers that are not only incomplete but potentially incorrect, as the model lacks the full context to resolve ambiguities or identify nuances.

RACE: Retrieval Augmented Context Engineering

To overcome these limitations, we propose a two-step approach that augments the initial retrieval with a generative processing stage. This architecture is designed to ensure that the final context provided to the generator is both comprehensive and compact.

Step 1: Broad Initial Retrieval

The process begins similarly to standard RAG, but with a key difference: the initial retrieval is intentionally broad. Instead of aiming for a small, highly relevant "top-k" set, we retrieve a much larger number of potentially relevant text snippets. The goal of this stage is recall over precision; we want to cast a wide net to ensure that all possible pieces of relevant information are captured, even if it means including a significant number of irrelevant documents.

Step 2: Generative Filtering and Summarization

This is the core of our proposed solution. The large set of retrieved documents is not passed directly to the final generator. Instead, each document is individually processed by an intermediate generative model (which can be the same as, or a smaller version of, the final generator). For each document, this model is tasked with one of two actions, conditioned on the original user query:

1. **Discard:** If the document is irrelevant to the user's question, the model is trained to generate a single, predefined token, such as "None". This is an extremely fast and efficient operation, allowing the system to quickly discard a large volume of irrelevant text in a single forward pass and with minimal computational overhead.
2. **Summarize:** If the document is relevant, the model does not pass the full text along. Instead, it generates a concise summary of the document. Crucially, this is not a generic summary; it is a **question-focused summary**. The model is prompted to extract and synthesize only the information within the document that is directly pertinent to answering the user's query. This process distills the document down to its essential, relevant components, removing extraneous information.

For example, if the query is about the market reaction to a report, and a retrieved document is a 50-page sustainability report, the summarizer would not summarize the entire report. It would extract only the key metrics and statements that the market would likely react to, resulting in a summary that is perhaps only a few sentences long but contains all the necessary information from that source.

Step 3: Final Generation with Synthesized Context

After the filtering and summarization step, we are left with a collection of short, dense, question-focused summaries. Because these summaries are significantly shorter than the original documents, they can now be concatenated, along with the original query, and fit comfortably within the context window of the final generator LLM.

The final generator is now equipped with a context that has been specifically curated for the task at hand. It contains all the relevant information, drawn from a wide range of initial sources, and it is presented in a compact, digestible format. This allows the model to perform the high-level synthesis and reasoning required to construct a comprehensive and nuanced answer, connecting the dots between the various pieces of summarized evidence.

Industrial Use Cases of RACE

The practical necessity for the RACE architecture is evident in complex challenges faced by industrial small and medium-sized enterprises (SMEs). Through our work on multiple TBI projects in the framework of the European Digital Innovation Hub (EDIH) AI5production, we have identified several recurring scenarios in both discrete manufacturing and process industries where standard RAG systems fall short. These use cases demonstrate a clear need for a more robust approach to information synthesis.

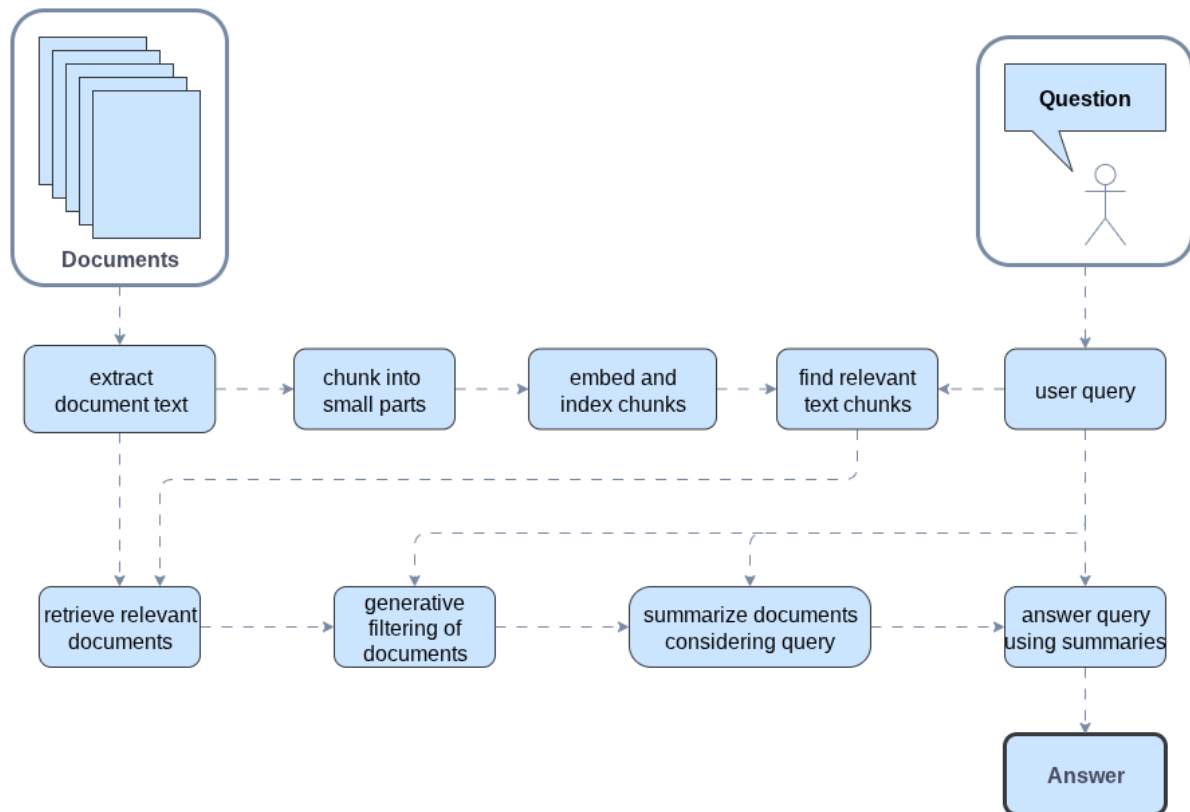


Figure 1. Retrieval Augmented Context Engineering Approach

These scenarios highlight a common industrial pattern: the need to make critical decisions by synthesizing incomplete, multi-source information. This is precisely where standard RAG fails and where the RACE framework provides a robust, practical solution.

Key industrial applications for RACE include:

- **Complex Product Formulation:** Developing a customer-specific chemical recipe, for instance, requires an expert to synthesize information from multiple technical manuals, safety data sheets, regulatory documents, and real-time material costs or availability. A standard RAG query might retrieve a base formula but miss a critical contraindication from a safety sheet or fail to suggest a cheaper, available alternative ingredient. RACE excels here by collecting all potentially relevant documents and providing concise, query-focused summaries, ensuring the final recommendation is safe, compliant, and cost-effective.
- **Dynamic Process Optimization:** To guarantee optimal plant configuration, an operator must consider specific material properties alongside equipment tolerances and process control parameters. A simple retrieval might not connect a new batch of raw material to the necessary adjustments in machine settings, leading to suboptimal output or equipment wear. RACE can synthesize data from material datasheets and operational manuals to suggest holistic adjustments, ensuring peak performance.

- **Resilient Supply Chain Management:** When a required component is out of stock, finding a viable substitute involves searching internal inventories, supplier catalogues, and external technical databases. Standard RAG struggles to bridge these disparate sources to find and validate an alternative. RACE's broad retrieval and summarization can identify potential substitutes and synthesize the necessary information, including required modifications to tooling or configurations, to solve the problem.
- **Automated Data Integration:** For manufacturers dealing with small, custom batch sizes, the manual extraction of data from technical sheets and order forms into production plans is slow and error prone. RACE can automate this workflow by extracting and summarizing key parameters from diverse source documents, feeding a complete and structured context into production planning systems.

Conclusion and Future Work

The standard RAG pipeline, while a significant step forward, is a brittle solution for complex information needs. It suffers from both imperfect retrieval and the hard constraint of context window size. The Retrieval Augmented Context Engineering (RACE) approach presented in this paper directly addresses these two fundamental weaknesses. By separating the initial broad retrieval from the final generation with an intelligent generative processing step, we can construct a near-optimal context for the generator. This process ensures that all relevant information is identified and preserved, while irrelevant noise is efficiently discarded.

The result is a system that achieves significantly higher quality in both its retrieval (by retaining all relevant information) and its generation (by providing a complete, synthesized context). This method allows the LLM to move beyond simple fact extraction and engage in the kind of complex synthesis and analysis that users increasingly demand.

Future work will focus on optimizing the generative summarization step, exploring the trade-offs between different model sizes for this intermediate task, and developing methods to create structured summaries (e.g., tables, JSON, code), rather than purely textual ones, to further enhance the final generator's reasoning capabilities.

Further on and in relation with the work presented in this paper, we will follow, investigate and contribute to the development of manufacturing foundation models and their industries'-specific finetuning, also in combination with new developments of the Model Context Protocol [5] that allows to build two-way (real-time) connections between AI systems and required data sources. Another interesting recent development related to the introduced work in this paper is called Agentic Retrieval-Augmented Generation (Agentic RAG) that enhances traditional RAG by incorporating autonomous AI agents that use planning, reflection, and collaboration to manage complex, multistep tasks and dynamically refine retrieval strategies [6]. By utilizing those recent developments in our own future work, we will actively be paving the way for further advancement of Generative AI solutions in industrial settings, with a specific focus on their utilization opportunities by industrial SMEs.

With the presented RACE approach and the envisaged further investigations and developments, we feel confident to be prepared for similar Tbl requests in the second EDIH phase as well as longer-term R&D collaborations with our SME partners.

References

- [1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [2] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," *arXiv preprint arXiv:2004.04906*, 2020. [Online]. Available: <https://arxiv.org/abs/2004.04906>

- [3] N. F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," arXiv preprint arXiv:2307.03172, 2023. [Online]. Available: <https://arxiv.org/abs/2307.03172>
- [4] Z. Yue et al., "Inference Scaling for Long-Context Retrieval Augmented Generation", Thirteenth International Conference on Learning Representations (ICLR 2025). [Online]. Available: <https://openreview.net/pdf?id=FSjIrOm1vz>
- [5] Anthropic, "Model Context Protocol," *Anthropic News*, Nov. 04, 2024. [Online]. Available: <https://www.anthropic.com/news/model-context-protocol>
- [6] A. Singh et al., "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG", arXiv preprint arXiv:2501.09136, 2025. [Online]. Available: <https://arxiv.org/abs/2501.09136>

Reconstruction pipeline for low-cost images for creating digital twins

Attila Szabo
VRVis GmbH
szabo@vrvis.at

Lisa Kellner
VRVis GmbH
kellner@vrvis.at

Martin Riegelneegg
VRVis GmbH
riegelneegg@vrvis.at

Georg Haaser
VRVis GmbH
haaser@vrvis.at

Christoph Traxler
VRVis GmbH
traxler@vrvis.at

Kurzzusammenfassung

Neuartige Technologien in den Bereichen LiDAR und Photogrammetrie halten zunehmend Einzug in die Baustellenüberwachung und versprechen eine schnellere sowie effizientere Vermessung. Ihr volles Potenzial bleibt aber ungenutzt - fehlende Softwareintegration und unausgereifte Verarbeitungswerkzeuge bremsen den praktischen Einsatz. In diesem Beitrag stellen wir eine modulare Pipeline zur Rekonstruktion digitaler Zwillinge vor, die im Rahmen von drei EDIH-Projekten entwickelt wurde. Ziel ist die effiziente Verarbeitung von Baustellendatensätzen, insbesondere solcher, die mittels Low-Cost Vermessungstechnologien erfasst wurden. Die Pipeline kombiniert moderne Verfahren aus der Computer Vision mit künstlicher Intelligenz zur Bereinigung von Punktwolken und zur automatisierten Geometrierekonstruktion.

Keywords: Digital Twin, Punktwolke, Computer Vision

Abstract

New developments in LiDAR and photogrammetry are rapidly finding their way into construction site monitoring, promising faster and more efficient surveying processes. Yet, the broader potential of these technologies is still constrained by limited software integration and the lack of mature processing tools. This contribution introduces a modular pipeline for creating digital twins, developed in the context of three EDIH projects. The aim is to streamline the handling of construction site data, with a particular focus on datasets collected using affordable surveying equipment. The approach combines modern computer vision methods with AI-driven point cloud cleaning and automated geometry reconstruction, providing a practical step toward more accessible and scalable digital twin generation.

Keywords: Digital Twin, Point Cloud, Computer Vision

Introduction

New technologies for low-cost measurement data collection are increasingly being used on construction sites and in infrastructure projects. In particular, end-user devices such as smartphones are commonly used as documentation tools, providing additional meta-information such as GPS location and time stamps. In addition, manufacturers of some laser scanners, such as XGRIDS [1], promise lower costs and time savings during recording.

These devices produce a large amount of data with varying quality characteristics. For example, modern smartphones come with a LiDAR component, but it only has a short range. Or point clouds from cheaper laser scanners contain more signal noise than those from more expensive models. Conventional point cloud processing pipelines are often ill-equipped to handle such cases. The lack of integrated tools and processes necessitates manual cleaning and tedious evaluation of the data sets.

We present an acquisition and analysis pipeline in several stages of the digital twin process. These projects, carried out as part of the EDIH program, enrich three successive steps in the digital twin pipeline with AI and computer vision algorithms.

Problem description

3D Construction Site - Efficient Scanning of Trenches with iPhone LiDAR

This article explores the potential of commercially available iPhones and their built-in depth sensors for construction site documentation.

This article includes documentation of a trench (a narrow ditch). The iPhone 15 Pro model used for surveying includes a LiDAR (Light Detection and Ranging) sensor with a comparatively low range and resolution. Recording and post-processing methods are explored to achieve acceptable accuracy when surveying the trench. The raw iPhone data is compared with reference scans taken at the same time with a professional TLS.

AI-supported automatic cleaning of point clouds

On construction sites, continuous documentation and progress monitoring using LiDAR or photogrammetry is desirable. However, new low-cost and high-speed technologies often result in point clouds that are significantly contaminated with unwanted error points (“noise points”). To counteract this problem, noise points must be removed from the point cloud. Traditionally, this is done in point cloud software, where lasso selections are repeatedly made manually from multiple angles to precisely select noise points in 3D. This is very labor-intensive and time-consuming.

In this article, two complementary techniques were developed to reduce the time required to clean up noise points. A simple artificial intelligence (AI)-assisted point selection is used instead of lasso selection. Furthermore, the cleaning process is semi-automated using point stability analysis from multiple scans.

RebarTwin: AI-accelerated geometric segmentation of reinforcement structures

Reinforcement structures are fundamental components of concrete structures. It is important to perform quality checks to ensure the structural integrity of the concrete structure. Quality control is traditionally performed manually. Inspectors must count large numbers of thin metal rods and compare their dimensions with construction plans. This process is time-consuming and prone to error.

In this article, the effort required for quality control of reinforcing steel is reduced through automated software support. The result is an end-to-end pipeline for the segmentation and reconstruction of cylindrical structures in point clouds, which can be measured and counted automatically.

Methodology

3D Construction Site - Efficient Scanning of Trenches with iPhone LiDAR

The survey took place at two locations, one in Bogenneusiedl and one in Groß-Enzersdorf. Three types of scan data were obtained:

- High-precision laser scan with a professional TLS.
- iPhone LiDAR raw data scan and processing in RTAB-Map [2].
- Photogrammetry using the iPhone camera.

A specially developed iOS app delivers all accessible sensor data from the iPhone Sensor API as a data stream. This data includes: internal/external camera parameters, color photos, depth image from the LiDAR sensor, segmentation mask, stable features, and camera trajectory data (GPS, compass).

Practical scanning methods are evaluated to identify best practices for scanning with an iPhone. This expands the current state-of-the-art, in which mainly specialized structures are reconstructed [5], but the general performance of this approach is rarely assessed.

AI-supported automatic cleaning of point clouds

One of the two cleaning methods offers simplified interactivity, while the other runs fully automatically. The techniques work synergistically, as each automatic process must ultimately be checked and corrected manually.

Traditional lasso selection, which requires many mouse clicks from multiple viewing angles, is replaced by semantic selection of objects with just one mouse click. A common AI image segmentation library is used for this purpose. The 3D points are projected onto a 2D depth image from a scanner position. The area of interest is centered around the mouse click point. The opening angle is calibrated so that the desired object occupies about half of the image, which has delivered the best results in this context. This technique was chosen because of its high performance and free availability, as the state-of-the-art 3D point segmentation approaches [6] are currently difficult to use and deliver.

Furthermore, point cloud cleaning is automated by analyzing point consistency. Most noise points are caused by movement or deformation of objects. Such inconsistencies between scans are detected by comparing individual observations. Real surfaces must appear consistent across multiple scans, which noise points do not. A consistency measure is calculated for each point based on three factors. Points that violate these criteria are noise points and are removed.

- The position in space in each scan
- The consistency of the reconstructed surface normals in each scan
- The number of scanners that match the observation

RebarTwin: AI-accelerated geometric segmentation of reinforcement structures

The technology consists of two components: (1) automatic isolation of reinforcement structures from a point cloud using image segmentation AI, and (2) reconstruction of cylindrical primitives. The data input is a photogrammetric point cloud that has already been pre-cleaned, for example using the approaches described above.

The first step is to automatically segment reinforcement structures from the larger point cloud. An AI image segmentation network recognizes reinforcing bars in photos. A synthetic data generation pipeline is used for training, eliminating the need for costly manual data annotation and allowing finer control over the training process.

After segmenting the reinforcement structures in the 3D point cloud, the individual metal bars are reconstructed as cylinder primitives. A region-growing algorithm tracks coherent cylindrical structures throughout the point cloud. Starting from a starting point, extrapolation and optimization are alternately applied to approximate the exact shape of the structure. In this use case, it is also important that the bends in the iron bars are explicitly modeled as torus segments. This approach corresponds approximately to the

State-of-the-Art [7], hebt sich aber durch die Einbettung in ein interaktives System von diesem ab.

Implementation and results

3D Construction Site - Efficient Scanning of Trenches with iPhone LiDAR

An ARSession is started on the iPhone via ARKit, which continuously records ARFrames. For each frame, photos, depth maps, and internal and external camera parameters can be accessed. Using the Open3D library, these data are used to generate a colored point cloud for each frame. The position and orientation of the camera can be used to align the individual frame point clouds with each other. This

creates a point cloud of the entire trench. Figure 1 below shows the photo and the corresponding depth map of a frame.

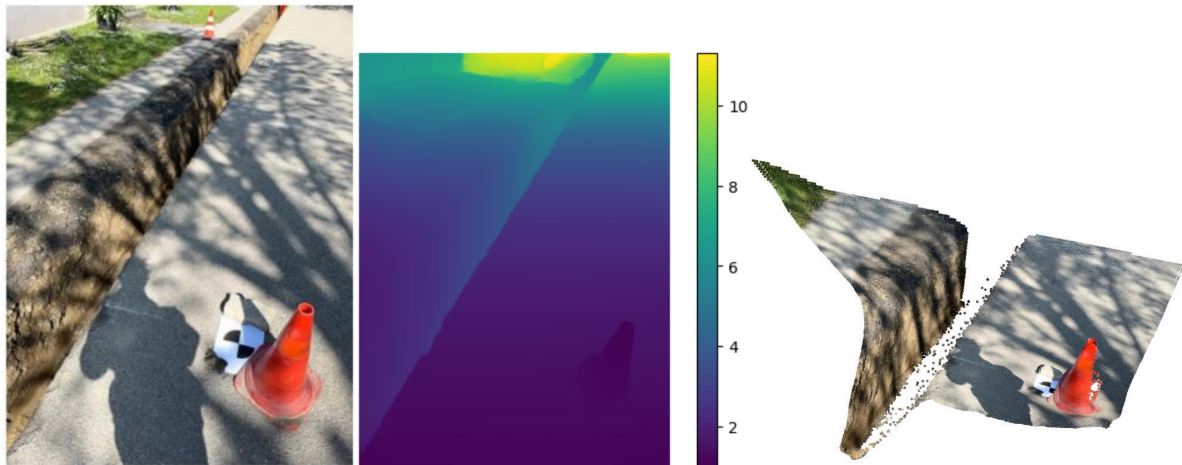


Figure 1. Photo taken with the iPhone (left), depth image (center), point cloud (right).

RTAB-Map is an open source solution for SLAM that supports loop closing using image-based features and markers. As a proof of concept, several Aruco markers were laid out and the scene was captured with the RTAB-Map iOS app.

Several practical recording methods were evaluated. Since one side of one of the gutters was difficult to access, both one-sided and two-sided inspections were used. The evaluation can be seen in Table 1 below.

Tabel 1. Comparison of different recording methods with the iPhone..

Recording type	Result
Portrait or landscape format	Landscape format captures more of the environment on either side of the trench.
Walking on one side only	One side of the trench is recorded more thinly.
Walking on one side, returning on the other	Both sides of the trench are close, but there is registration drift.
Selfie stick	Provides the best result, as you can hold the iPhone in the trench and walking on one side is sufficient.

Figure 2 below shows a cross-section of various scans, each registered to the reference point cloud (blue). The scans are registered with the iPhone LiDAR (“1st + 2nd half”), unregistered (‘Original’), the commercial scanning app “RTAB-Map,” and the photogrammetric reconstruction from “COLMAP.”

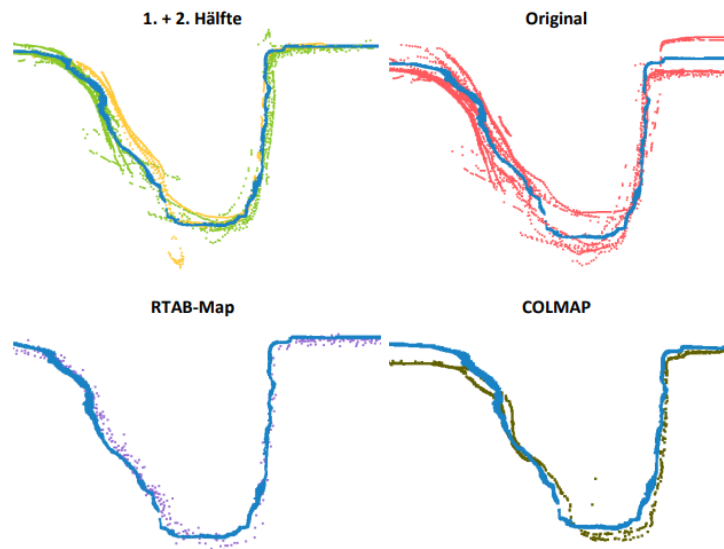


Figure 2. Comparison of images captured via iPhone LiDAR (top left), unregistered (top right), scan app “RTAB-Map” (bottom left), and photogrammetry (bottom right), each relative to the reference point cloud (blue)).

In summary, it can be said that the iPhone Pro (from generation 12 onwards) with its sensors is suitable for the 3D construction site in the case of the trench. The depth data from the iPhone and the point clouds reconstructed from it showed in both visual and numerical comparisons that they deviate by only a few centimeters from the reference scan, which should be sufficient in the case of the trench.

AI-supported automatic cleaning of point clouds

The freely available segmentation library SegmentAnything [3] is used for AI point cloud cleaning. Using ONNX [4], the AI weights were exported from the Torch+CUDA framework to a manufacturer-independent environment. This allows segmentation to be performed on customer hardware without an NVidia graphics card. An example of a person selected in this way can be seen in Figure 3 below.

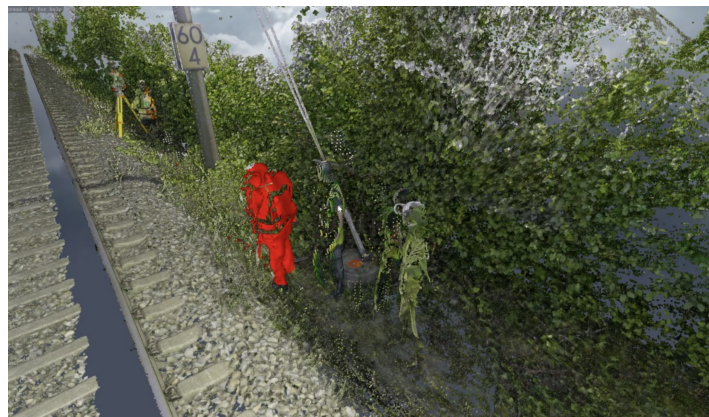


Figure 3. Person selected in a point cloud using AI methods.

For automatic point cloud cleaning, a point is located in all panoramas, the normals are reconstructed, and a corresponding decision criterion is applied. In Figure 4 below, the quality measure for each point is visualized as a heat map. It can be seen that the automatic selection (top) corresponds to the manual cleaning (bottom) in essential aspects.

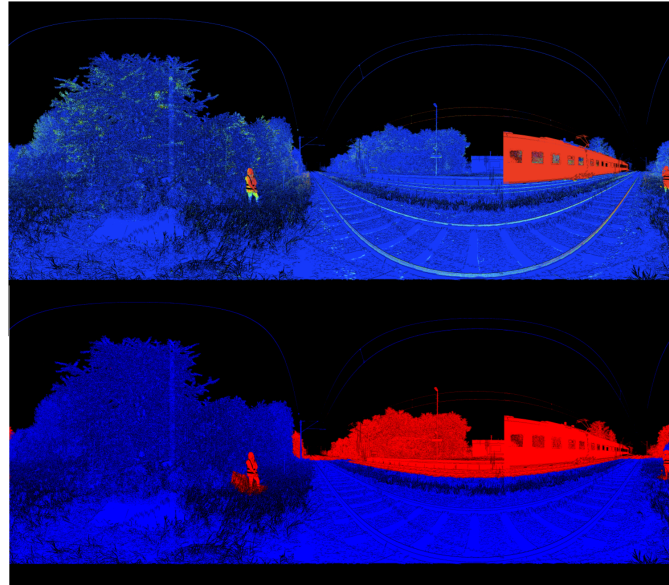


Figure 4. Top: People and train selected in the scan using automatic point cloud analysis. Bottom: Manual point cloud cleaning..

In summary, it can be observed that the point cloud cleaning techniques presented can significantly speed up the traditional approach. In some cases, the cleaning can even be made more accurate, for example, if too much has been selected when drawing the lasso manually.

RebarTwin: AI-accelerated geometric segmentation of reinforcement structures

Deep learning networks require a large amount of annotated data for training. For this purpose, reinforcement bars are recreated in Blender, randomly placed in the scene, and provided with background images. Examples are shown in Figure 5 below.

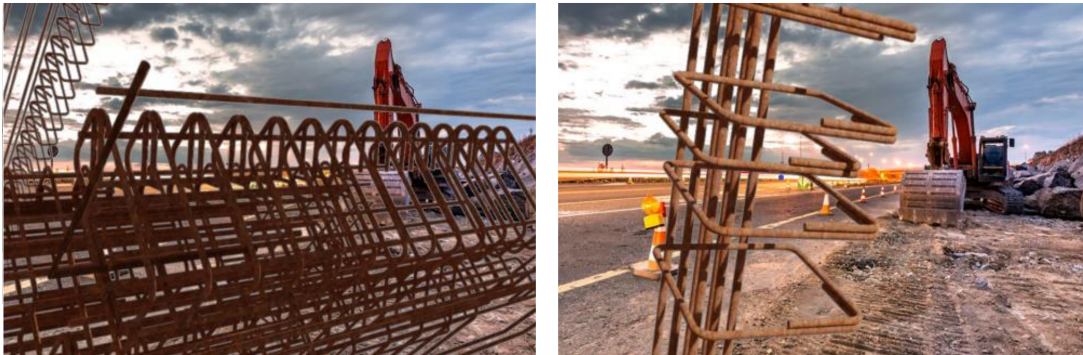


Figure 5. Synthetically generated photos of reinforcing bars for training AI segmentation.

A Feature Pyramid Network (FPN) with a Mixed Transformer (MiT) as its backbone was chosen as the network architecture. With FPN, feature maps of various sizes are generated with little additional effort, enabling the network to recognize objects at different scales and classify them as reinforcement or non-reinforcement. The network achieved good segmentation performance with real test photos, as can be seen in Figure 6 below.

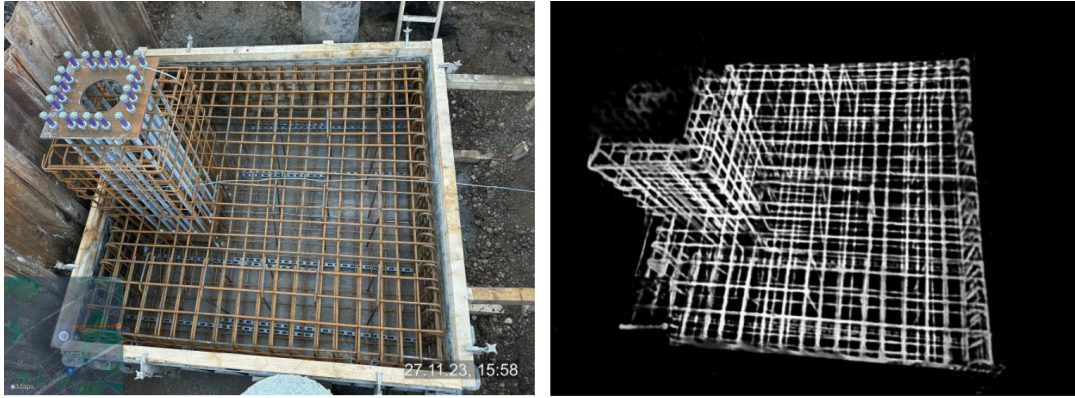


Figure 6. Left: Photo of reinforcement cage. Right: AI-segmented mask..

The reconstruction of the cylinders already works well in the early prototype stage. In the sample data set, the reconstruction error of a rod was approximately 0.8 mm with a diameter of 12 mm. The scaling of the point cloud was not entirely accurate (scaled with SLAM). The different bar dimensions could be clearly distinguished. An example reconstruction can be seen in Figure 7 below.

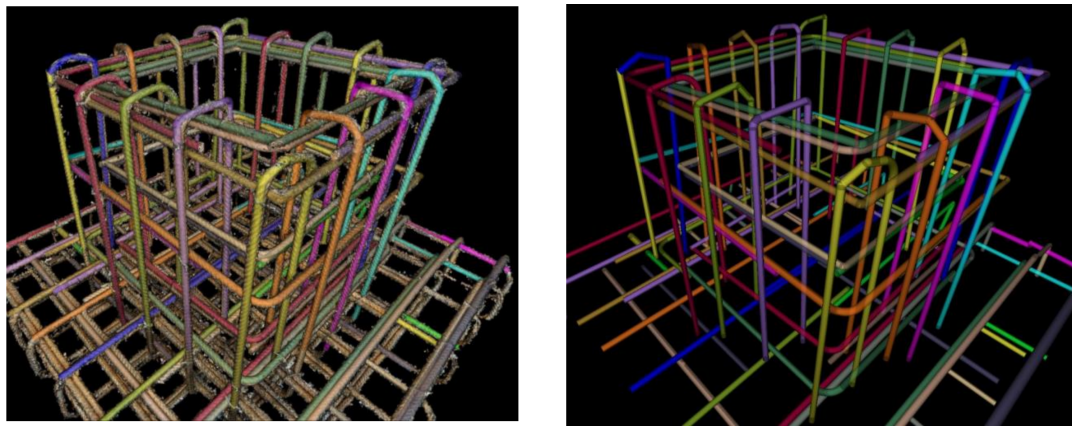


Figure 7. Reinforcement bars reconstructed from the point cloud using the region growing method.

Summary

This article presented a recording and reconstruction pipeline using low-cost devices. Since the individual projects within the EDIH program were intended as proof-of-concept, they already show promising possibilities at this stage. Conventional consumer devices are suitable for use as recording devices for complex data with higher quality requirements when used with appropriate software processing. The techniques presented should therefore be particularly useful for industries that work with digital twins and carry out regular inventory or inspections.

The approaches presented in this article can be transferred to other industries to varying degrees:

iPhone LiDAR for data acquisition: The use of commercially available smartphones enables high-quality 3D recordings by non-specialized professionals and is therefore suitable for trades such as installation, masonry, or electrical work.

Cleaning up noisy point clouds: Data from uncontrolled environments or low-cost sensors is often distorted by movement or reflections. Automated cleaning offers significant advantages here, for example in mobile mapping.

AI-supported image segmentation: Integration into surveying software is already commonplace and allows image data to be evaluated for a wide range of applications, including construction site monitoring, industrial manufacturing, and medical imaging.

References

- [1] XGRIDS. "XGRIDS-3D Reconstruction-Spatial Computing-Handheld Scanner." Xgrids.com, 2025, www.xgrids.com/?page=geomatics. Accessed 29 July 2025.
- [2] IntRoLab. "RTAB-Map." RTAB-Map, introlab.github.io/rtabmap/. Accessed 29 July 2025.
- [3] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
- [4] Developers, Onnx Runtime. ONNX Runtime. 2021, <https://onnxruntime.ai/>.
- [5] Mizutani, Tsukasa, and Shunsuke Iwai. "Improving construction site efficiency through automated progress monitoring of underground pipe installation sites using image color analysis of iPhone LiDAR camera data." *Developments in the Built Environment* 20 (2024): 100557.
- [6] Zhou, Yuchen, et al. 'Point-SAM: Promptable 3D Segmentation Model for Point Clouds'. *arXiv [Cs.CV]*, 2024, arxiv.org/abs/2406.17741. arXiv.
- [7] Qiu, Rongqi, Qian-Yi Zhou, and Ulrich Neumann. "Pipe-run extraction and reconstruction from point clouds." *European Conference on Computer Vision*. Cham: Springer International Publishing, 2014.

Towards Optical Tracking for AR Assisted Oral Surgery

Philipp Fleck
VRVis GmbH
philipp.fleck@vrvis.at

Clemens Arth
VRVis GmbH
clemens.arth@vrvis.at

Kurzzusammenfassung

Jüngste Entwicklungen im Bereich Augmented Reality (AR) ermöglichen eine Vielzahl bisher ungeahnter Anwendungsgebiete und -bereiche. Insbesondere unterstützt AR-Technologie medizinische Aufgaben rund um den menschlichen Körper. In diesem Beitrag wenden wir AR-Technologie im zahnmedizinischen Bereich an, wo Zahnärzte oder Ärzte Mundhöhle und Zähne untersuchen. Wir beschreiben die ersten Schritte zur Posenschätzung einer Kamera anhand eines 3D-Modells von Teilen der Mundhöhle. Dabei nutzen wir Deep Learning und überlagern die Umrisse der oberen Gierung mit einem Live-Kamerabild. Wir identifizieren zwei Gruppen von Ansätzen: modellbasiertes, merkmalsloses Tracking und deren medizinische Anwendungen. Wir konzentrieren uns detaillierter auf den letztgenannten Ansatz und konzentrieren uns auf merkmalsbasiertes und merkmalsloses Tracking. Der erste Schritt basiert auf einem einfachen Kantenerkennungsalgorithmus für das Modell und das tatsächliche Livebild. Der zweite Schritt besteht darin, ein dreidimensionales (3D-)Bild der Zähne mit dem Livebild zu überlagern. Neben dem medizinischen Bereich ist unser Ansatz auch auf andere texturlose starre Objekte anwendbar, bei denen Konturbilder extrahiert werden können.

Keywords: model-based tracking, dental, feature-less tracking, augmented reality, oral surgery

Abstract

Recent developments in augmented reality (AR) allow for a wide range of never-thought-of applications and domains. In particular, AR technology allows for the support of medical tasks around the human body. In this work, we apply AR technology to the dental domain, where a dentist or physician investigates the oral cavity and teeth. We describe the first steps towards pose estimation of a camera with respect to a 3D model of parts of the oral cavities, leveraging deep learning and overlaying the outlines of the upper jaw onto a live camera feed. We identify two groups of approaches, namely, model-based, feature-less tracking and their medical applications. We focus on the latter type of approach in more detail, and we focus on feature-based tracking and featureless tracking. The first step is based on a simple edge detection algorithm on the model and the actual live image. The second step is to overlay a three-dimensional (3D) image of the teeth onto the live image feed. Besides the medical domain, our approach is applicable to other texture-less rigid objects where contour images are extractable.

Keywords: model-based tracking, dental, feature-less tracking, augmented reality, oral surgery

Introduction

Recent developments in augmented reality (AR) allow for a wide range of never-thought-of applications and domains. In this work, we focus on a human-centered use case to support medical tasks around the human body. In particular, we apply AR technology to the dental domain, where a dentist or physician investigates the oral cavity and teeth. This is of high interest to the industry and scientific community, as many "off-the-book tasks" require extensive preparation. Such tasks include the preparation of implants, which sometimes involve the construction of a custom-made drilling jig for each patient. Furthermore, the result of such procedures is hardly presentable to patients since there is no "on-patient" visualization using AR or similar technologies.

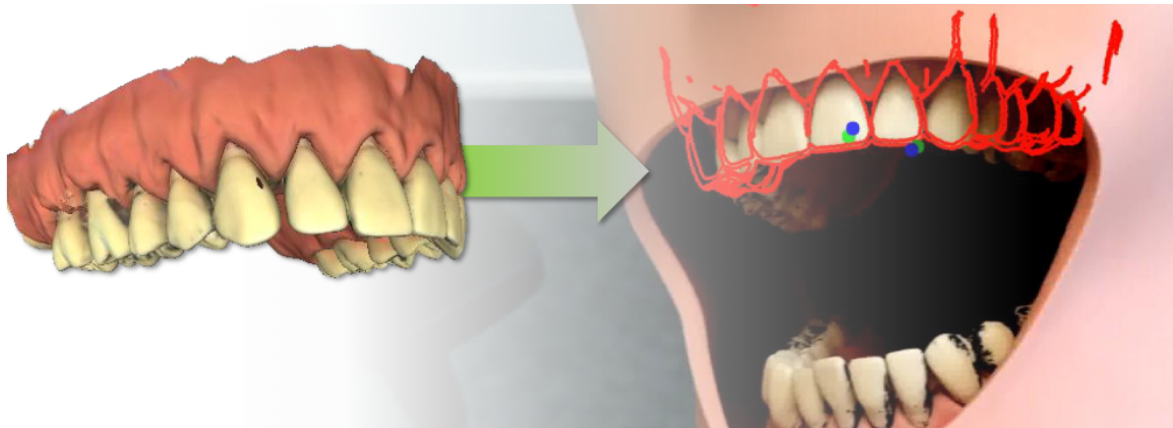


Figure 1. Conceptual drawing of the proposed approach. On the left, a 3D model of the upper jaw is shown. As a result, on the right the teeth outlines are overlaid on an image during live mode.

The ability to spatially address an individual tooth or other feature in the oral cavity enables a new range of applications. Ranging from pre-flight training, over in-situ visualization, to navigational support during operation, we can only guess the added value. However, we can still draw conclusions from other similar fields with manual hand operation. With the rise of Industry 4.0 and the digitization trend, the manufacturing and maintenance of machinery created the need for more state-of-the-art guidance. In this regard, systems that support operators performing a task were found to be very beneficial. The use of AR visualizations in situ was thus manifested to empower workers in different domains.

Our aim is now to translate these findings into the dental domain and to identify a possible future roadmap towards oral surgery that uses AR support. We therefore describe the first steps towards pose estimation of a camera with respect to a 3D model of parts of the oral cavity, leveraging deep learning and overlaying the outlines of the upper jaw onto a live camera feed (see Figure 1 for an exemplary picture).

Related Work

The first steps in AR already demonstrated the potential to precisely track and enhance the perceived world [1], [2]. However, due to technical limitations, markers [3], [4], [5] were the first choice to robustly fulfill the core requirement for AR, namely the 6-degree-of-freedom pose tracking, i.e. robustly estimating the geometric relationship between a camera, respectively, display, and the world at frame rate. Although marker tracking is still common due to its robustness especially in challenging industrial environments, it is nowadays considered old-fashioned, and placing markers is also impractical in certain use cases. Therefore, over the last 15 years research has gradually shifted towards markerless tracking. In that regard, we can identify two groups of approaches, namely a) feature-based tracking and b) feature-less tracking. Since teeth have few distinctive features, we focus on the latter type of approach in more detail, i.e. on model-based, feature-less approaches and their medical applications.

Model based tracking

In contrast to marker- or natural feature-based approaches, model-based methods try to estimate the pose based on the given shape, i.e. a geometrical model. Such techniques are especially interesting for textureless rigid objects. Rehg et. al. [6] demonstrated how to overcome self-occlusion with model-based tracking already back in 1995. Comport et. al. [7] introduced a robust model-based version for robot vision. Wüst et. al. [8] utilized lines in a multi-hypothesis procedure to estimate the pose of the objects. Later, with the support of CAD models as source model, Wüst et. al. [9] were able to track industrial models, but also estimate the initial pose [10] in a more robust way. Usually such approaches are based on lines, hollows, depth, or shape and try to minimize the distance between the source representation (i.e. the model) and the observation.

Iterative closest point algorithms [11], voting algorithms [12] or similar are used to establish correspondences through matching, which then serve as input to pose estimation. With the development of powerful handheld devices, real-time image processing tracking of CD models [13] on mobiles became possible. Seo et. al. [14] present a direct method based on monocular RGB images. More recently, MOLTR [15] was capable of localizing, tracking, and reconstructing multiple objects from live daily scenes. Model-based tracking also finds its application in the automotive or autonomous driving field, where sparse LIDAR data are used to track pedestrians [16].

Our presented approach takes inspiration from the approaches of the last 20 years and applies them to the dental setting, in which camera tracking for AR applications has not been widely evaluated or established.

Medical AR

Within the medical domain, assistive systems [17] became more and more important in recent years. Since medical personnel often undergoes varying stress levels [18] it is beneficial to reduce mental loads [19] for certain tasks such as navigation and to free up brain capacity for tasks such as problem solving. This problem is observed in all medical disciplines, but also in dental surgery [20].

Wang et. al. [21] presented an AR based system using a halfway mirror setup to create visual overlays for the operator on the stationary patient. An example was to highlight an individual tooth within the oral cavity. Other approaches, such as Karner et. al. [22] or Weber et. al. [23] propose complete head tracking based on CT-Scan or instant data, which are hardly applicable to dental use cases. One particular aspect in dentistry are deformations caused by instruments such as mouth openers or similar devices. The oral cavity is mainly composed of soft and flexible tissue, so the assumptions about rigidity that are the core of most tracking approaches hold only for bones and teeth at most.

Dental reconstruction [24] and facial reconstruction [25] are used for preparation and real-time navigation. Other methods [26] used reconstructions to create precise implants for structural defects. In the dental industry, intraoral scanners and jaw tracking systems are increasing in popularity [27].

However, our work strives to close the gap between well-known tracking algorithms in AR and special conditions in the dental domain, by creating new algorithms adopted to the human characteristic and the connected application cases.

Methodology

Traditional model-based tracking approaches consist of two major parts: a) the model creation and b) tracking. The acquisition step takes care of creating the trackable model. Depending on the features and technology used for the tracking, the model has to be prepared accordingly. For example, as Wüst et. al. [13] presented, the lines and other information are extracted from the CAD data to be used at runtime. Such algorithms require a known initialization pose of the model since there are only limited possibilities. Either the search space becomes large and extensive search has to be performed to initialize the tracking algorithm, or more commonly, the user aligns a predefined initial pose with the real world and reduces the search space by a magnitude.

Model Creation

Our assumption is the existence of an oral cavity scan, which is available as a 3D model (Figure 1, left) for further tracking purposes. Because such scans are very precise and up to scale, the model can directly serve as the basis for further data extraction and preparation stages for the subsequent tracking algorithm.

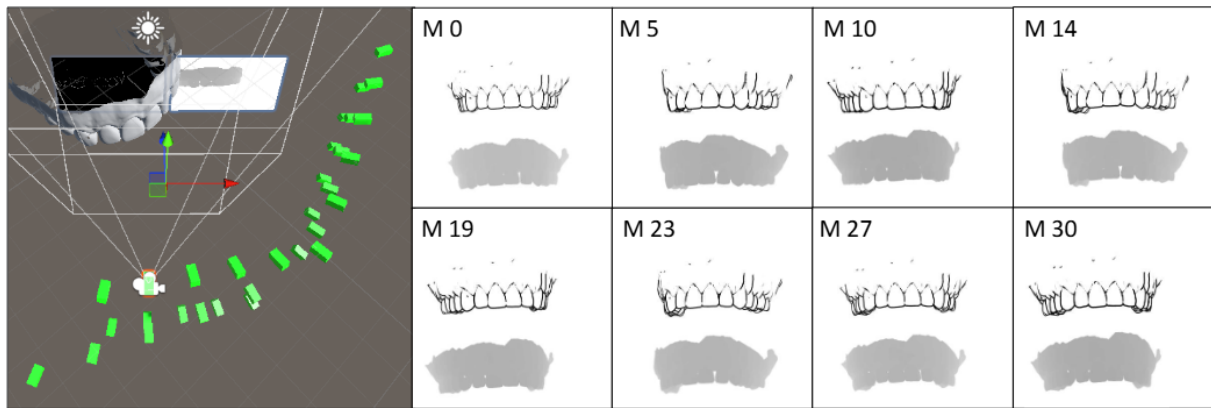


Figure 2. The 3D reconstruction of the upper jaw is loaded in Unity (left) and poses are sampled from different orientations (green). The resulting silhouette images and depth renderings are shown on the right "M 0" related to the first model pose.

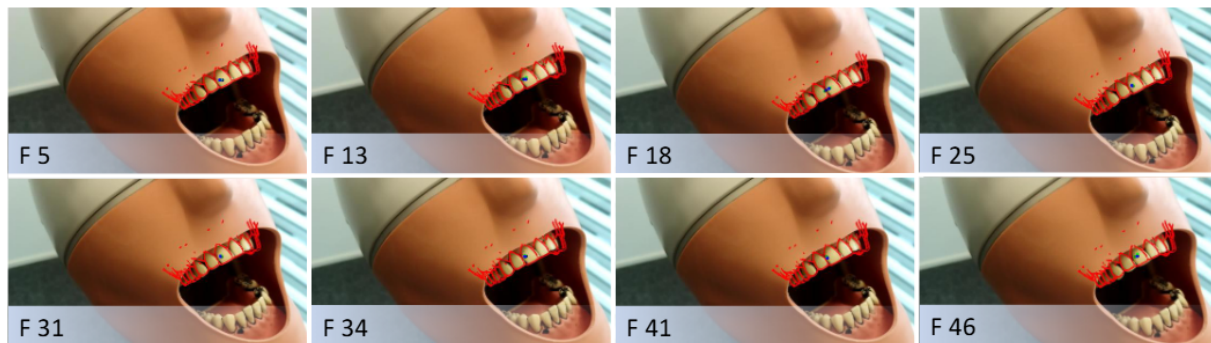


Figure 3. Tracking result over selection of frames of an 80 Frame long sequence. The prediction stays on target, and tries to optimize its alignment each frame.

We import the scan into a game engine, i.e. Unity, to create a variety of 2D images. The scan consists of the upper and lower jaw, the teeth, and the surrounding soft tissue. The images generated mimic the appearance of parts of the oral cavity as seen from a camera during tracking; therefore, we can extract any set of distinctive features that are known from long-time CV research; however, the major issue here is the absence of such features.

We therefore opt for an approach based on deep learning. We focus on the upper jaw for the moment and create a large number of artificial images which capture the appearance within a given hemispheric space. Although we do not use depth information, we also create a set of depth images alongside the 2D images of the model as seen from a known pose. As described in the work of Rad and Lepetit [28], we train a model to infer the pose based on a newly seen image. Some examples of the extracted images are shown in Figure 2. Note that training can be done in a very fast way because of the overall low number of images and the fact that we only need to estimate the pose of a single model (as opposed to an entire set of different objects).

Tracking

The tracking is implicitly solved by the network; i.e. the inference of the network is actually a pose of the object with respect to the camera position. The network is also able to infer intermediate poses, which means that the discrete space of samples and poses used for training becomes a continuous space for tracking; however, because the network is still trained from a defined number of samples, we add an optimization step for the pose, as we assume that the camera is handheld during tracking.

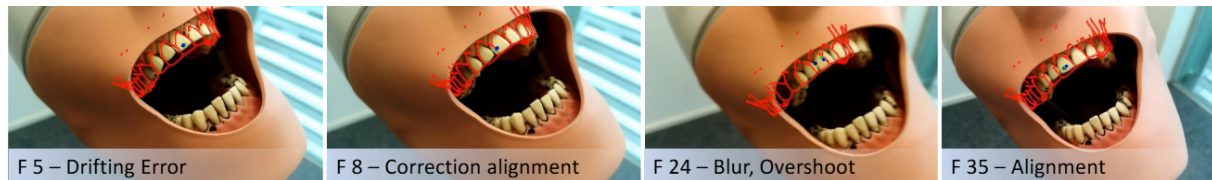


Figure 4. Selected frames out of a 50 frames sequence showing alternating error-alignment occurrences either due to drift or overshoot.

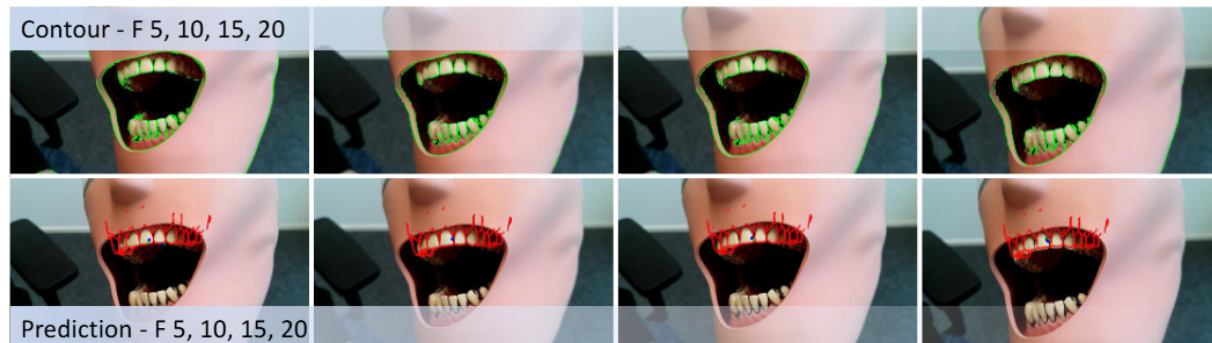


Figure 5. Visual comparison of the extracted input frame contour (green) against the predicted model (red). The limited light conditions within the mouth cavity cause a reduced amount of contours towards the back side of the jaw leading to inaccurate predictions in frame 15 and 20 (red).

We use a simplistic Kalman filter to model the camera motion. The pose representation, although based on only 6 degrees of freedom, is usually given as a 3×4 matrix, encapsulating a 3×3 matrix for the rotation and a 1×3 column vector for the translation. This representation cannot be easily used within an optimization approach. We therefore transform the pose into the $SE3$ Lie space, which forms a differentiable manifold. The Kalman filter smoothes the motion path between previous and new inferences. For visualization, the resulting pose is transformed back into its original format.

Visualization

To assess the overall fit, we use a simplistic edge detection algorithm on the model and the actual live image. This enables us to render the most prominent contours from our model and visually compare those with the contours in the live image. This is facilitated by importing the camera images into Unity and rendering the model contours into the images using a custom shader.

Experiments

We conducted several experiments using a dummy model of a patient's head. The 3D model of the upper jaw was created and processed as previously described. The overall development of our prototype is at an early stage; therefore, we only show a sequence of frames from a single test video. Also note that only the upper jaw was considered so far, and tracking the lower jaw (which may essentially be subject to a different type of motion compared to the rest of the patient's head) is considered future work.

In Figure 3 we show a selection of frames out of an 80-frame test sequence. Our prediction is capable of following the movement, but also re-adjusting just when the error increases. In such cases, the algorithm tries to minimize the distance to the observation by running the alignment step in a larger area. As with many prediction algorithms, such as Extended Kalman Filters, edge cases such as abrupt changes in direction can cause "overshooting", which is then compensated for the frame after occurrences as depicted in Figure 4.

In Figure 5, we compare the contour extracted from the input stream with the predicted model. Looking closely at frame 15 and 20, we identify a shorter contour towards the inside of the mouth cavity, due to lighting and blurriness. However, this causes the prediction to become inaccurate in the lower distribution in the upper jaw. In a real-world scenario, a dentist would ensure good lighting conditions,

which further helps reduce blur. In the future, we plan to gather more test data (also of human participants) for a more in-depth evaluation under real-world conditions.

Discussion and Conclusion

In this work, we described the first steps towards optical tracking of parts of the oral cavity based on deep learning. Besides the medical domain, other industrial applications, such as, part tracking, training, assembly guidance, are also meaningful domains. Similarly, VisionLib [13] focuses on CAD-based data in manufacturing for AR-based tracking applications. Furthermore, AR use-cases [29] demonstrated recently there benefit when enhancing the real-world with additional useful data.

We roughly outlined our initial approach and showed a couple of experimental results. Despite being a valid approach to the given problem, we found several deficiencies in our current prototypical implementation. First, the framework used for the initial evaluation is no longer considered state-of-the-art and was also originally built for a different use case. In deep learning research, it is important to strip the design and ultimate functionality of a neural network to its core requirements, both to save computational power and to increase accuracy at the same time. Therefore, we consider the application of more recent concepts [30] as the next step forward.

A second deficiency was found in the use of autofocus cameras, which are regularly used today without much reflection. However, with respect to the use case at hand, a system should ultimately provide both precise registration and accurate video overlay. This can hardly be achieved in setups where introspection or control of the camera used for live capture is not possible, i.e. the varying focal length of the camera is known in the subsequent visualization procedure. We therefore consider the use of cameras with a fixed focus to be a better choice and contrary to what we used in our previous experiments.

Finally, the system is not applicable in real time. Future efforts concern the use of more lightweight hardware such that a practical implementation is possible without large-scale GPU power or even server resources.

Acknowledgment

This work was funded by the European Digital Innovation Hub (EDIH) under contract No. 101083472. However, views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

- [1] S. Henderson und S. Feiner, „Exploring the benefits of augmented reality documentation for maintenance and repair“, *IEEE transactions on visualization and computer graphics*, Bd. 17, Nr. 10, S. 1355–1368, 2010.
- [2] S. Gauglitz, B. Nuernberger, M. Turk, und T. Höllerer, „World-stabilized annotations and virtual scene navigation for remote collaboration“, in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, S. 449–459.
- [3] H. Kato, „Inside ARToolKit“, in *1st IEEE international workshop on augmented reality toolkit*, 2007.
- [4] D. Wagner und D. Schmalstieg, „Artoolkitplus for pose tracking on mobile devices“, 2007.
- [5] M. Fiala, „ARTag, a fiducial marker system using digital techniques“, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, S. 590–596 Bd. 2. doi: 10.1109/CVPR.2005.74.
- [6] J. M. Rehg und T. Kanade, „Model-based tracking of self-occluding articulated objects“, in *Proceedings of IEEE International Conference on Computer Vision*, 1995, S. 612–617. doi: 10.1109/ICCV.1995.466882.
- [7] A. I. Comport, É. Marchand, und F. Chaumette, „Robust model-based tracking for robot vision“, in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, IEEE, 2004, S. 692–697.
- [8] Wuest, H. and Vial, F. and Stricker, D, „Adaptive line tracking with multiple hypotheses for augmented reality“, in *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05)*, IEEE, 2005, S. 62–69.

- [9] H. Wuest und D. Stricker, „Tracking of industrial objects by using cad models“, *Journal of Virtual Reality and Broadcasting*, Bd. 4, 2007.
- [10] H. Wuest, F. Wientapper, und D. Stricker, „Adaptable model-based tracking using analysis-by-synthesis techniques“, in *Int. Conference on Computer Analysis of Images and Patterns*, Springer, 2007, S. 20–27.
- [11] D. Chetverikov, D. Svirko, D. Stepanov, und P. Krsek, „The trimmed iterative closest point algorithm“, in *2002 International Conference on Pattern Recognition*, IEEE, 2002, S. 545–548.
- [12] D. Chang und S. Hashimoto, „An inverse voting algorithm for Hough transform“, in *Proceedings of 1st International Conference on Image Processing*, IEEE, 1994, S. 223–227.
- [13] H. Wuest, T. Engelke, F. Wientapper, F. Schmitt, und J. Keil, „From CAD to 3D Tracking — Enhancing & Scaling Model-Based Tracking for Industrial Appliances“, in *IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2016, S. 346–347. doi: 10.1109/ISMAR-Adjunct.2016.0114.
- [14] B.-K. Seo und H. Wuest, „A direct method for robust model-based 3D object tracking from a monocular RGB image“, in *Computer Vision—ECCV 2016 Workshops, Proc., Part III 14*, Springer, 2016, S. 551–562.
- [15] K. Li, H. Rezatofighi, und I. Reid, „MOLTR: Multiple Object Localization, Tracking and Reconstruction From Monocular RGB Videos“, *IEEE Robotics and Automation Letters*, Bd. 6, Nr. 2, S. 3341–3348, 2021, doi: 10.1109/LRA.2021.3061080.
- [16] P. Held, D. Steinhauser, A. Koch, T. Brandmeier, und U. T. Schwarz, „A Novel Approach for Model-Based Pedestrian Tracking Using Automotive Radar“, *IEEE Transactions on Intelligent Transportation Systems*, Bd. 23, Nr. 7, S. 7082–7095, 2022, doi: 10.1109/TITS.2021.3066680.
- [17] R. Nagyné Elek und T. Haidegger, „Non-technical skill assessment and mental load evaluation in robot-assisted minimally invasive surgery“, *Sensors*, Bd. 21, Nr. 8, S. 2666, 2021.
- [18] D. Yu u. a., „Quantifying intraoperative workloads across the surgical team roles: room for better balance?“, *World journal of surgery*, Bd. 40, S. 1565–1574, 2016.
- [19] A. Plessas, R. Witton, M. Paisi, F. Ellwood, und E. McColl, „Using the Mental Health Wellness Framework in dental practice“, *Bdj Team*, Bd. 9, Nr. 10, S. 10–12, 2022.
- [20] M. Maleki Roveshti, M. Raci, und F. Valipour, „Examining Mental Workload and Incidence of Musculoskeletal Abnormalities of Dentists During Surgery“, *Archives of Rehabilitation*, Bd. 25, Nr. 2, S. 232–247, 2024.
- [21] J. Wang u. a., „Augmented Reality Navigation With Automatic Marker-Free Image Registration Using 3-D Image Overlay for Dental Surgery“, *IEEE Transactions on Biomedical Engineering*, Bd. 61, Nr. 4, S. 1295–1304, 2014, doi: 10.1109/TBME.2014.2301191.
- [22] F. Karner u. a., „Single-Shot Deep Volumetric Regression for Mobile Medical Augmented Reality“, in *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, Cham: Springer International Publishing, Jan. 2020, S. 64–74.
- [23] M. Weber, D. Wild, J. Kleesiek, J. Egger, und C. Gsaxner, „Deep Learning-Based Point Cloud Registration for Augmented Reality-Guided Surgery“, in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, S. 1–5. doi: 10.1109/ISBI56570.2024.10635333.
- [24] I.-B. Pavaloiu, N. Goga, I. Marin, und A. Vasilateanu, „Automatic segmentation for 3D dental reconstruction“, in *2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2015, S. 1–6.
- [25] W. Zielonka, T. Bolkart, und J. Thies, „Towards metrical reconstruction of human faces“, in *European conference on computer vision*, Springer, 2022, S. 250–269.
- [26] J. Li u. a., „Back to the roots: Reconstructing large and complex cranial defects using an image-based statistical shape model“, *Journal of Medical Systems*, Bd. 48, Nr. 1, S. 55, 2024.
- [27] M. Revilla-León, D. E. Kois, J. M. Zeitler, W. Att, und J. C. Kois, „An overview of the digital occlusion technologies: Intraoral scanners, jaw tracking systems, and computerized occlusal analysis devices“, *Journal of Esthetic and Restorative Dentistry*, Bd. 35, Nr. 5, S. 735–744, 2023.
- [28] M. Rad und V. Lepetit, „BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth“, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, IEEE Computer Society, 2017, S. 3848–3856. doi: 10.1109/ICCV.2017.413.
- [29] A. Stanescu u. a., „Error Management for Augmented Reality Assembly Instructions“, in *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2024, S. 690–699. doi: 10.1109/ISMAR62088.2024.00084.
- [30] B. Wen, W. Yang, J. Kautz, und S. Birchfield, „FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects“, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Juni 2024, S. 17868–17879. doi: 10.1109/CVPR52733.2024.01692.

14:30

EINLASS & GET TOGETHER

Austausch mit Expert:innen, Poster & Live Use Cases

15:00

BEGRÜSSUNG

- Rektor Univ.-Prof. Dr. Stefan Koch (JKU Linz)
- Rektorat TU Wien
- Vertreter:in der Europäischen Kommission

15:15

**KEYNOTE - STRATEGIEN FÜR
UNTERNEHMEN IM ZEITALTER VON KI UND INNOVATION**

15:45

LÖSUNGSANSÄTZE IN DER PRODUKTION (I)

- *Innovationsimpulse zur Digitalisierung von KMUs*
Barbara Krumay & Manuel Mühlburger, JKU Linz
- *Künstliche Intelligenz in der Produktion*
Franz Sümec, TU Wien & Barbara Krumay, JKU Linz
- *RAG is Not All You Need! - Echtzeitinformationen im Kontext mit KI verarbeiten*
Raphael Pisoni & Mario Pichler, Software Competence Center Hagenberg GmbH
- *Digitale mensch-zentrierte Assistenzsysteme*,
Harald Bauer, Gerhard Ebenhofer, Christoph Heindl & Gernot Stübl, PROFACTOR GmbH

16:45

PAUSE

17:00

LÖSUNGSANSÄTZE IN DER PRODUKTION (II)

- *Rekonstruktionspipeline für Low-Cost Aufnahmen zur Erstellung von Digital Twins*
Georg Haaser, Lisa Kellner, Martin Riegelneegg, Attila Szabo & Christoph Traxler, VRVIs
- *Open Source Cybersecurity Anomalieerkennung*,
Andreas Ekelhart, Alexander Szönyi & Markus Klemen, SBA Research GmbH
- *Intelligente Chatbots zur Abfrage von Unternehmenswissen - ALLEX*
Lukas Pichlmann, LKR Leichtmetallkompetenzzentrum Ranshofen
- *Benefits für produzierende Unternehmen durch das Europäische Innovationsökosystem*
Roland Sommer & Stefanie Werderits, Plattform Industrie 4.0

18:00

VON AI5PRODUCTION ZU AI5INNOVATION

WEITERE DREI JAHRE ALS AI5INNOVATION BIS 2028

- *Fortsetzung von AI5production als AI5Innovation bis 2028*
Franz Sümec, TU Wien, Coordinator AI5Innovation
- *Bundesminister Dr. Wolfgang Hattmannsdorfer
Gemeinderat Almir Balihodzic & weitere Vertreter:innen aus Politik und Wirtschaft
zur Bedeutung der künstlichen Intelligenz für Österreich in wirtschaftlich
herausfordernden Zeiten, der Stärkung des Technologiestandortes Linz durch
gezielte Innovationsinitiativen, der Rolle der EDIHs der European Digital Innovation
Hubs für Unternehmen und das Land Oberösterreich und die Europäische
Ausrichtung der Künstlichen Intelligenz*

18:30

NETWORKING MIT BUFFET & LIVE-MUSIK



AI5 INNOVATION

Start 11/2025



Co-funded by
the European Union



European
Digital Innovation
Hubs Network



Bundesministerium
Wirtschaft, Energie
und Tourismus



FFG
Forschung wirkt.