Karlsruher Schriften
zur Anthropomatik

Band 65

Jürgen Beyerer, Tim Zander (Eds.)

Proceedings of the 2023 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

KIT Scientific Publishing

Jürgen Beyerer, Tim Zander (Eds.)

Proceedings of the 2023 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

Karlsruher Schriften zur Anthropomatik
Band 65
Herausgeber: Prof. Dr.-Ing. habil. Jürgen Beyerer

Eine Übersicht aller bisher in dieser Schriftenreihe
erschienenen Bände finden Sie am Ende des Buchs.

Proceedings of the 2023 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

by
Jürgen Beyerer, Tim Zander (Eds.)

KIT Scientific Publishing

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark
of Karlsruhe Institute of Technology.
Reprint using the book cover is not allowed.

www.ksp.kit.edu

Print on Demand 2024 – Gedruckt auf FSC-zertifiziertem Papier

ISSN 1863-6489 (Schriftenreihe)
ISSN 2510-7259 (Tagungsband)
ISBN 978-3-7315-1351-3
DOI 10.5445/KSP/1000168973

Preface

In 2023, the annual joint workshop of the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB) and the Vision and Fusion Laboratory (IES) of the Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT) was hosted again in a Black Forest house near Triberg.

For a week from the 30th of July to the 4th of August, the PhD students of both institutions delivered extended reports on the status of their research and participated in heated discussions on topics ranging from computer vision, industrial production, control theory to large language models. Most results and ideas presented at the workshop are collected in this book in the form of detailed technical reports. This volume provides a comprehensive and up-to-date overview of some of the research programs of the IES Laboratory and the Fraunhofer IOSB.

The editors thank Jonas Vogl and Zeyun Zhong for their efforts resulting in a pleasant and inspiring atmosphere throughout the week. We would also like to thank the doctoral students for writing and reviewing the technical reports as well as for responding to the comments and suggestions of their colleagues.

Jürgen Beyerer & Tim Zander

Contents

Formal Process Maturity Measure

Negar Arabizadeh

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
negar.arabizadeh@kit.edu

Abstract

Measuring the maturity of a process instance is essential because we can evaluate the progress toward a mature production process; Consider we have added some sensors or actuators in the manufacturing process so we need a formal basement for measuring the maturity of the process after these changes. We used the definitions in the control theory to provide a formal measure for process maturity. We defined *Elucidability* E, *Forcability* F, and *Supervisability* S that are ostensive, interpretable, and based on quantities that can be determined or estimated. These lead to a formal definition for a measure of the process maturity M, that combines technical and economic considerations.

1 Introduction

Consider we have a manufacturing process and we do some software or hardware changes to it. 'maturity' generally can be defined as 'the state of being complete, perfect or ready'[8]. It is valuable for us to know how much we were successful to have progress in achieving our matured process[6][7]. P_i^j shows the process instance with 'hardware revision', i and 'software revision', j. The transition $P_i^j \rightarrow P_i^{j+1}$ shows software changes like control optimization and the $P_i^j \rightarrow P_{i+1}^j$ shows hardware changes and structural modifications. During

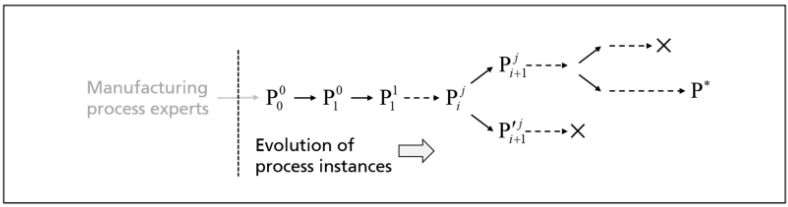


Figure 1.1: Flow diagram of the maturation of a manufacturing process. The symbol $\times$ indicates abortion, and P^* denotes the desired final mature process instance.

these changes in the project, we will have an evaluation chain of the project instance. Some changes might be done parallel in the project process and they compete with each other. In Figure 1.1 which shows the process evaluation chains, $\times$ shows abortion which means at the end of a chain we did not achieve the desired maturity, and P^* shows we achieved the desired matured process instance. Figure 1.2 shows a process instance. The input signals x and the output signals y are defined to be observable, and all non-observable dynamic quantities are in s. If the added sensor or actuator instrumentation allows components of s to be observable, those quantities are assigned to be additional components of y and x of the next process instance P_i^j. The noise n might happen in the system and the system has controller π.

The paper is organized as follows, in Section 2 we will mention the Controllability and Observability definitions. In Section 3 we explain our new definitions which are Forcability, Elucidability, and Supervisability, their formulas, and their relations with definitions which are in Section 2. In Section 4, we have a simulation example of an inverted pendulum on which we implemented our method. In the end, the paper is concluded in Section 5.

2 Controllability and Observability Definitions

Here we have the definitions of controllability and observability[4].

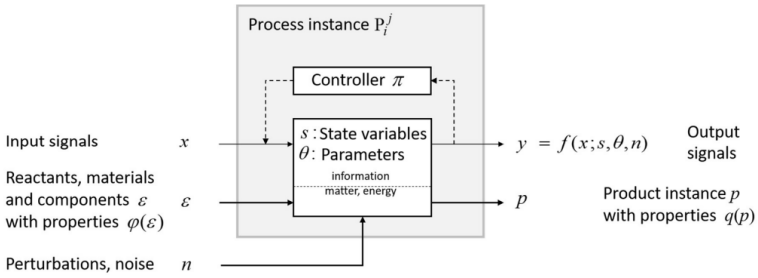


Figure 1.2: Abstract view of a process instance. This may represent the complete production process or a sub-process at any hierarchical process level. The variables x, y, n, and s are time-dependent signals.

Controllability

The system states s is controllable if there exists a control signal $\pi(t)$ over a finite time interval $0 \leq t \leq t_1$ such that a target state s^* is reached at t_1. When a system state is controllable, makes it possible to control this state to the desired point.

Observability

The system state s is observable if the state $s(0)$ at time $t = 0$ can be fully determined by the observations $y(t)$ during a finite time interval $t_0 \leq t \leq 0$.

In these definitions of controllability and observability, we can say the system is controllable/observable or not; Also it does not consider any uncertainty in the system. In this paper, we will introduce new definitions related to observability and controllability that gives us a degree of controllability and observability and considers the uncertainty in the system.

In ML, the concept of PAC learning (probably approximately correct) [9] describes the number of data samples minimally required to achieve a model error smaller than some η with a probability larger than $1 - \delta$. We apply similar notions to introduce definitions.

3 Definitions for measuring the maturity

Elucidability

For the process instance, P_i^j Elucidability which is similar to observability is approached by defining the minimum observation time τ that is necessary to achieve a certain estimation precision for the state s and parameters θ. So we find the probability of having the estimation error less than a constant precision η_θ and η_s. The internal states of the process can be estimated from the time series X_T, Y_T during the finite observation in duration T, so $X_T := \{x(0), ..., x(T)\}$, $Y_T := \{y(0), ..., y(T)\}$. The formula for Elucidability is

$$\begin{aligned} &E(T, \eta_s, \eta_\theta) := \\ &Pr(\|s(T) - \hat{s}(X_T, Y_T)\| < \eta_s \wedge \|\theta - \hat{\theta}(X_T, Y_T)\| < \eta_\theta | s(0)) \end{aligned} \tag{3.1}$$

In which the $\hat{s}(X_T, Y_T)$ is the estimation of state $s(T)$ at time T, with time series X_T and Y_T in duration T; It is the same for $\hat{\theta}(X_T, Y_T)$ which is the estimation of θ. If we do not have the time duration T we can find it with the smallest time to achieve a specific error η_s and η_θ, with a probability larger than $1 - \delta$,

$$\tau(\eta_s, \eta_\theta, \delta) := min\{T > 0 | E(T, \eta_s, \eta_\theta) > 1 - \delta\}. \tag{3.2}$$

Forcability

Forcability F which is similar to controllability is to be defined to measure the capability to steer the values of y and s with x toward target values $\breve{y}$ and $\breve{s}$. The most effective way to steer the process is to establish a closed-loop control to react instantaneously to the dynamic answers of the process. To quantify Forcability of a process instance with a given controller π, the probability of being able to force the process after a time T into the neighborhood of target values $\breve{y}$ and $\breve{s}$ is defined as follows,

$$\begin{aligned} &F_\pi(T, \eta_s, \eta_y) := \\ &Pr(\|y(T) - \breve{y}\| < \eta_y \wedge \|s(T) - \breve{s}\| < \eta_s \mid s(0), x(t) = \pi(Y_t), \breve{y}, \breve{s}) \end{aligned} \tag{3.3}$$

In close loop control, X_T is determined from the control policy π. In a condition that we do not have the $y(T)$ and $s(T)$ and we need to estimate them, we can replace the estimated $\hat{y}(T)$ and $\hat{s}(T)$ with the real value of them, so the modified formula for Forcability is

$$\begin{aligned} &F_\pi(T, \eta_s, \eta_y) := \\ &Pr(\|\hat{y}(T) - \breve{y}\| < \eta_y \wedge \|\hat{s}(T) - \breve{s}\| < \eta_s \,|\, s(0), x(t) = \pi(Y_t), \breve{y}, \breve{s}) \end{aligned} \tag{3.4}$$

If we implement an optimal controller our formula will be changed to

$$F(T, \eta_s, \eta_y) := F_{\pi^*}(T, \eta_s, \eta_y) \tag{3.5}$$

If the time duration T is not fixed for a given process instance, we can use the PAC [9] in a way to find the minimum time we need to have a specific error η_s and η_y with the probability larger than $1 - \delta$,

$$\tau(\eta_s, \eta_y, \delta) = min\{T > 0 | F(T, \eta_s, \eta_y) > 1 - \delta\}. \tag{3.6}$$

Supervisability

Supervisability S is defined to measure the maturity of the quality of the process. In [1] there is information about quality management. The formula is written to find the probability of having the quality in the desired set Q in considering we have set our controller π and the input signal x, the state s, and parameter θ are in their desired set. The formula for Supervisability is defined as:

$$S(P_i^j) := Pr(q \in Q | x \in X_{adm}, s \in S_{adm}, \theta \in \Theta_{adm}). \tag{3.7}$$

In which $X_{adm}, S_{adm}, \Theta_{adm}$ are our desired set for x, s, θ. S measures the maturity of a process only from the technical point of view. To find out whether the resulting optimized process P^* is economic, the cost of the process also must be considered. Applying the economic part is beyond the scope of this project but it can be considered in the future.

4 Example

To clarify the definitions we have run an example in Matlab. The example is for an Inverted Pendulum [2] which is linearized around its stable state. The system has four states s_1, s_2, s_3 and s_4; We consider that we have added a sensor for observing the state s_3 in the matured system.

Figure 4.1 compares the probability of having a special error at different times for the state s_1.

Figure 4.2 compares the probability of having a special error at different times for the state s_2.

Figure 4.3 compares the probability of having a special error at different times for the state s_3. It is obvious that there is a comparable difference between the not Matured system and the matured system for the state s_3 for which we have added the sensor. For the not matured system the probability of having the observation error smaller than a specific constant number until some time is zero and after that growth, but for the matured system the probability started to grow at an earlier time.

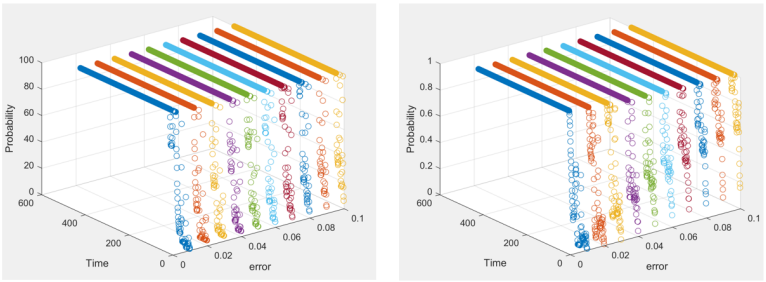
Figure 4.4 compares the probability of having a special error at different times for the state s_4. Also, this is obvious that for the state s_4 for the matured system, the probability of having an error less than a special constant number started to grow earlier.

Figure 4.5 shows the probability of having an error less than the special number for state s_1, this figure shows the probability of the smallest time that this error happens.

Figure 4.6 shows the probability of having an error less than the special number for state s_2, this figure shows the probability of the smallest time that this error happens.

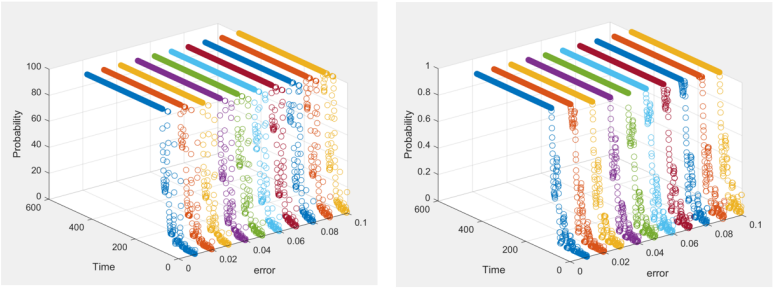
Figure 4.7 shows the probability of having an error less than the special number for state s_3, this figure shows the probability of the smallest time that the specific error happens. It is obvious that for the matured system the smallest time shifted to the earlier time.

Figure 4.8 shows the probability of having an error less than the special number for state s_4, this figure shows the probability of the smallest time that the specific error happens. It is obvious that for the matured system the smallest time shifted to the earlier time.



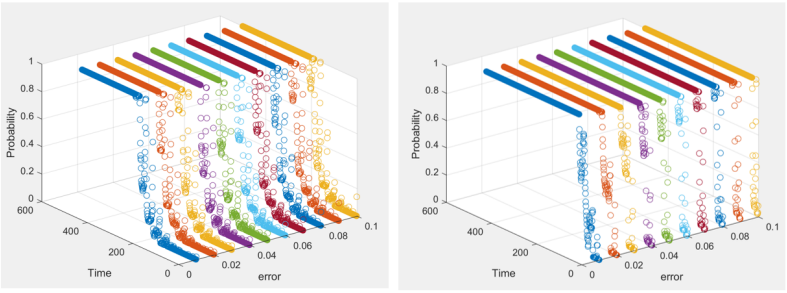
Not Matured system Matured system

Figure 4.1: Comparing the probability of having error less than a constant number at different times, for the state s_1; Left: For the not matured system, Right: For the matured system



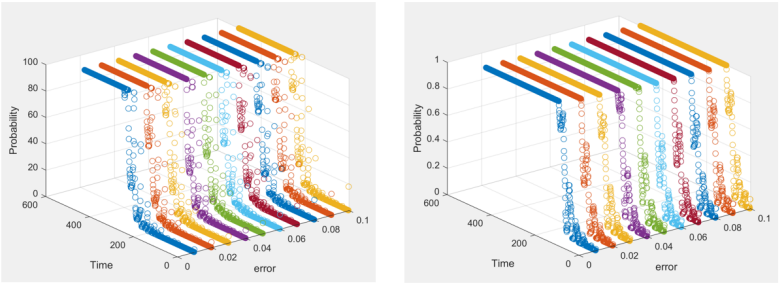
Not matured system Matured system

Figure 4.2: Comparing the probability of having an error less than a constant number at different times for the state s_2; Left: For the not matured system, Right: For the matured system



Not matured system Matured system

Figure 4.3: Comparing the probability of having an error less than a constant number at different times for the state s_3; Left: For the not matured system, Right: For the matured system



Not matured system Matured system

Figure 4.4: Comparing the probability of having an error less than a constant number at different times for the state s_4; Left: For the not matured system, Right: For the matured system

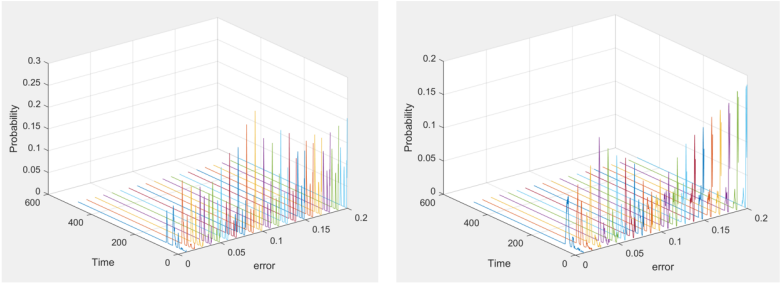


Figure 4.5: Comparing the probability of having an error less than a constant number at different times for the state s_1, this figure shows the smallest time that this error occurred; Left: for the not matured system Right: for the matured system

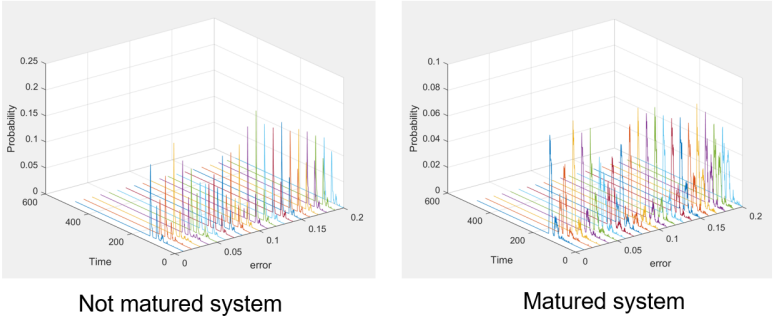
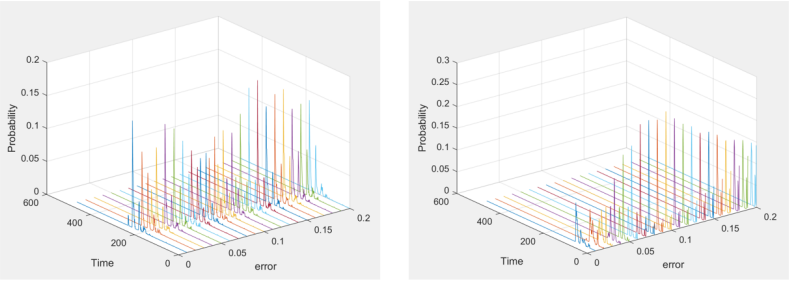
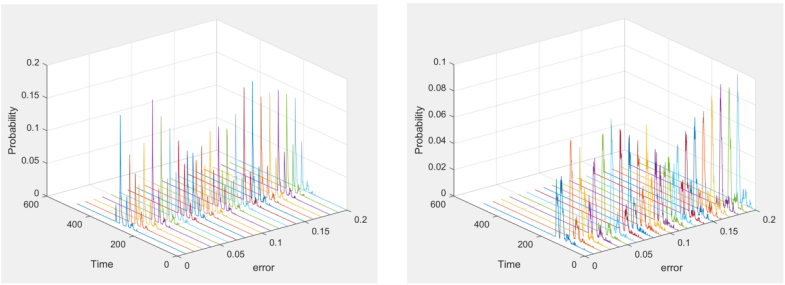


Figure 4.6: Comparing the probability of having an error less than a constant number at different times for the state s_2, this figure shows the smallest time that this error occurred; Left: for the not matured system Right: for the matured system



Not matured system Matured system

Figure 4.7: Comparing the probability of having an error less than a constant number at different times for the state s_3, this figure shows the smallest time that this error occurred; Left: for the not matured system Right: for the matured system



Not matured system Matured system

Figure 4.8: Comparing the probability of having an error less than a constant number at different times for the state s_4, this figure shows the smallest time that this error occurred; Left: for the not matured system Right: for the matured system

5 Conclusion

In this paper, we introduce definitions for measuring the maturity of a process. The Elucidability E which is similar to observability in the control theory is used for measuring the capability in estimating states and parameters of the system. Forcability F which is similar to controllability in control theory is used for measuring the capability to steer the states and outputs of the system toward their desired value. Supervisability S is defined for measuring the quality of the process. We run an example to show maturity in a process. In the example, we knew the exact physical model of the system; In practice, the exact model of the system might not be available but we can receive knowledge about the physical model of the system from experiments and have a partial model of the system and use it in machine learning[10][5][3].

References

[1] John C Anderson, Manus Rungtusanatham, and Roger G Schroeder. "A theory of quality management underlying the Deming management method". In: *Academy of management Review* 19.3 (1994), pp. 472–509.

[2] Olfa Boubaker. "The inverted pendulum: A fundamental benchmark in control theory and robotics". In: *International conference on education and e-learning innovations*. IEEE. 2012, pp. 1–6.

[3] Ricky TQ Chen et al. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).

[4] Rudolf E Kalman. "On the general theory of control systems". In: *Proceedings First International Conference on Automatic Control, Moscow, USSR*. 1960, pp. 481–492.

[5] George Em Karniadakis et al. "Physics-informed machine learning". In: *Nature Reviews Physics* 3.6 (2021), pp. 422–440.

[6] Tobias Mettler. "Maturity assessment models: a design science research approach". In: *International Journal of Society Systems Science* 3.1-2 (2011), pp. 81–98.

[7] Steven Peters. “A readiness level model for new manufacturing technologies”. In: *Production Engineering* 9 (2015), pp. 647–654.

[8] Catherine Soanes, Angus Stevenson, et al. *Concise oxford English dictionary*. Vol. 11. Oxford university press Oxford, 2004.

[9] Leslie G Valiant. “A theory of the learnable”. In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[10] Jared Willard et al. “Integrating physics-based modeling with machine learning: A survey”. In: *arXiv preprint arXiv:2003.04919* 1.1 (2020), pp. 1–34.

Securing XAI through Trusted Computing

Maximilian Becker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
maximilian.becker@kit.edu

Abstract

The escalating use of Artificial Intelligence (AI) and Machine Learning (ML) systems underscores the need for transparency and data security. This paper explores the fusion of Explainable AI (XAI) with trusted computing technologies such as Trusted Platform Modules (TPMs) and Trusted Execution Environments (TEEs). Highlighting the synergy between XAI, aimed at elucidating ML decision-making, and trusted computing, which fortifies system integrity, this study introduces novel approaches. Specifically, it proposes leveraging TEEs to protect user privacy during XAI computation and TPMs to verify system trustworthiness. This integration seeks to augment trust in AI systems by securing personal data processing and ensuring system integrity, thereby potentially reshaping the landscape of trust in AI technologies.

1 Introduction

AI and ML have become ever more utilized in recent years. This also increases the demand for XAI to increase the transparency of these systems. Because such systems are used more frequently they also process more and more personal data. The European Union is planning on putting forward certain transparency requirements for high-risk applications of AI [4]. This would increase the

demand for XAI even more. It order to convince users to share their personal data with AI systems they need to trust the systems. Trust is often mentioned in the XAI literature [9]. It is hard to trust a system you do not understand; XAI can help in this regard by making the AI system more understandable. However, this trust is not warranted if the system processing the personal data can be tempered with. To avoid this trusted computing technologies can be used. This combination could ensure trust in the system through trusted computing and trust in the AI through XAI.

We present some background on XAI and trusted computing in Section 2. Afterwards in Section 3 we show some concepts on how XAI and trusted computing can be combined. Section 4 ends with a short summary.

2 Background

Here we present some background, first on XAI and later about trusted computing technologies.

2.1 XAI

The goal of XAI is to make ML systems more transparent by generating explanations [2]. These explanations can explain individual decisions of ML models or the models as a whole. There are many different techniques utilized in XAI and the explanations themselves can also take different forms. The explanations can show different features and how important they are for the model, they can show alternatives that can change a prediction or show correlations between different features. Different explanation methods can be used for different use cases and user groups. Some explanation methods are very technical and produce complicated graphs. These methods can help developers to gain deep insight into the models they train. There are also methods that are oriented more towards end users. Such methods can for example deliver explanations in natural language understandable without technical knowledge.

The European Union is working on the Artificial Intelligence Act (AI Act) [4], a legislation that puts forward transparency requirements for high-risk applications

of AI. This regulation would require XAI to fulfill these requirements. The AI Act specifically mentions that "Users should be able to interpret the system output and use it appropriately". This focus on users of XAI systems would increase the demand for user-centered explanations.

2.2 Trusted Computing

The goal of trusted computing is to create trust in the hardware of a system as well as the software running on that system. The Trusted Computing Group (TCG)[1] creates standards around trusted computing. Most notably they created the specifications for the Trusted Platform Module (TPM)[2].

A TPM is a trusted hardware component build into many modern systems [5]. TPMs can add IT-security functionalities to a system. The main purpose of TPMs revolves around encryption. They can generate cryptographic keys and store generated keys as well as keys provided to the TPM. These keys can encrypt and decrypt data provided to the TPM. Additionally, the TPM can encrypt keys with a Storage Root Key to store them outside the TPM.

A TPM can be used to verify that a system is in a trusted software state [11]. To achieve this it has several Platform Configuration Registers (PCRs) which can be used to store a hash of the system state. Because the number of registers is limited new states are concatenated to the current hash, hashed again and the register gets overwritten. TPMs also have a functionality called enhanced authorization. Keys can be bound to policies and can only be used if e.g. expected values are stored in the PCRs. Alternatively the PCR values can be provided to users or applications to verify the system state from outside.

The process of verifying the state of a system from outside is called remote attestation. According to Coker et. al. "Remote attestation is the activity of making a claim about properties of a target by supplying evidence to an appraiser over a network." [3]. This process involves two parties: The attester provides its current state and the verifier wants to verify this state [11]. There are two

[1] https://trustedcomputinggroup.org/
[2] https://trustedcomputinggroup.org/resource/tpm-library-specification/

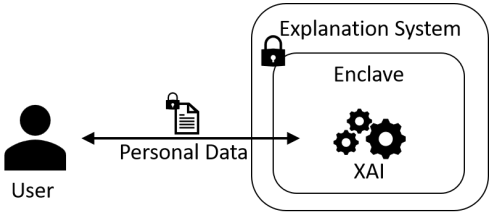
approaches to remote attestation: explicit attestation and implicit attestation. For explicit attestation with TPMs the attester provides the verifier with different values signed by the TPM including the PCR values from the TPM. The verifier checks these values by comparing them to know, trusted values. The attestation also involves a so called nonce that the verifier sends the attester at the beginning of the process to avoid replay attacks. Implicit attestation uses the enhanced authorization features of the TPM. The attester has to communicate with the verifier with a key that can only be used if the attester is in a known, trusted state. The key is bound to a policy an can therefore only be used if the system is in a trusted state. Through the key the verifier can verify that the attester can be trusted. At the beginning the verifier also sends a nonce that is then signed by the TPM and send back to verify that the system state is current.

Another trusted computing technology are Trusted Execution Environments (TEEs) [10]. A TEE is an isolated part of a processor with encrypted memory. It can be used to create enclaves in which data can be processed confidentially without other processes, even with higher privileges, being able to access the data. Examples for this technology are ARM TrustZone [1] and Intel Software Guard Extensions (Intel SGX) [6]. A similar technology is Intel Trust Domain Extensions (Intel TDX) [7] which builds on SGX. The idea there is to secure a complete virtual machine from its host. This means that for example a cloud provider can not access the VM but everyone that has access to it has to be included into the trust model.

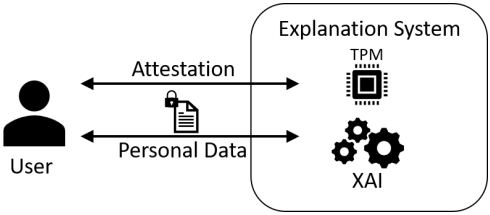
3 Explainable AI and Trusted Computing

XAI literature often mentions that XAI can be used to increase trust in AI systems [9]. However XAI is not sufficient to trust the system from an IT-security standpoint. To achieve this trusted computing can be used. There are different trusted computing technologies that are used in different contexts and give different guarantees. Trusted Execution Environments (TEEs) can create secured enclaves that are protected from the rest of the system. Trusted Platform Modules (TPMs) can be used to verify that a system is in a known, trustworthy state. Both technologies could be used to protect XAI applications.

To our knowledge there are no existing approaches that combine XAI and trusted computing. Section 3.1 presents a concept for TEEs and Section 3.2 a concept for TPMs. Figure 5.1 shows the two concepts for combining XAI and trusted computing.



(a) Calculating explanations in an enclave.



(b) Attesting the system state with a TPM.

Figure 3.1: Two concepts for combining XAI and trusted computing.

3.1 Securing XAI with TEEs

There are already some works that combine machine learning and trusted computing. One method that could be applied to XAI is origami inference proposed by Narra et. al. [8]. The proposed method is used to do privacy-preserving inference of deep neural networks using Intel SGX. First the encrypted user data that should be protected is received in the SGX enclave. The inference of the first layers of the deep neural network are calculated in a privacy-preserving way using SGX. Because the input is almost impossible to reconstruct after a

few layers, later layers can be executed as usually. This is done because TEEs have limited storage and processing capabilities. The input data is kept safe but execution time is greatly reduced compared to calculating everything in the enclave. XAI approaches also need to process personal data if the explanations are designated to end users. The methods need to evaluate the ML model, often multiple times to generate explanations. Origami inference could be used here to preserve the users privacy while maintaining an acceptable processing time. Figure 3.1(a) shows the concept of calculating explanations in enclaves without origami inference.

3.2 Securing XAI with TPMs

Another way to secure user data when generating XAI explanations is the use of TPMs. Remote attestation explained in section 2.2 can be used to verify that the system calculating the explanations is in a trustworthy state. Explicit as well as implicit attestation can be used to achieve this. This would require an attestation step to be executed before the use of the XAI application which could however be completely transparent to the user. An application such as an XAI dashboard could execute implicit attestation and the user would only be able to connect to the dashboard when the attestation succeeds. Figure 3.1(b) shows the concept of attesting the state of an explanation system with a TPM.

4 Summary

We presented two concepts for creating more trust in XAI systems by using trusted computing technologies. The first concept utilizes TEEs while the second one uses TPMs. These technologies could be especially interesting when the scope of the system involves processing personal data. In such scenarios the utilization of trusted computing could increase the users' trust in the system and make them more comfortable sharing personal data with the system.

References

[1] *ARM Security Technology Building a Secure System using TrustZone Technology*. 2008. URL: https://documentation-service.arm.com/static/5f212796500e883ab8e74531?token=.

[2] Nadia Burkart and Marco F Huber. "A survey on the explainability of supervised machine learning". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[3] George Coker et al. "Principles of remote attestation". In: *International Journal of Information Security* 10 (2011), pp. 63–81.

[4] European Commission. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[5] *Enterprise Security: Putting the TPM to Work*. 2008. URL: https://trustedcomputinggroup.org/wp-content/uploads/TPM-Applications-Whitepaper.pdf.

[6] *Intel Software Guard Extensions*. URL: https://www.intel.com/content/www/us/en/architecture-and-technology/software-guard-extensions.html.

[7] *Intel Trust Domain Extensions*. 2021. URL: https://cdrdv2-public.intel.com/690419/TDX-Whitepaper-February2022.pdf.

[8] Krishna Giri Narra et al. "Origami inference: private inference using hardware enclaves". In: *2021 IEEE 14th International Conference on Cloud Computing (CLOUD)*. IEEE. 2021, pp. 78–84.

[9] Atul Rawal et al. "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives". In: *IEEE Transactions on Artificial Intelligence* 3.6 (2021), pp. 852–866.

[10] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. “Trusted execution environment: what it is, and what it is not”. In: *2015 IEEE Trustcom/BigDataSE/Ispa*. Vol. 1. IEEE. 2015, pp. 57–64.

[11] *TCG Trusted Attestation Protocol (TAP) Information Model for TPM Families 1.2 and 2.0 and DICE Family 1.0*. 2019. URL: https://trustedcomputinggroup.org/wp-content/uploads/TNC_TAP_Information_Model_v1.00_r0.36-FINAL.pdf.

Causal Representation Learning: A Quick Survey

Frank Doehner

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
frank.doehner@kit.edu

Abstract

Causal representation learning (CRL) has recently become an object of intensive research. Representation learning aims to infer a lower dimensional, but meaningful, representation from a given set of data, effectively increasing interpretability and processability. Disentangled representation learning applies an independency constraint onto the inferred latent variables of the representation. The applicability of such frameworks on real world data is limited, as absolute independence between all generating factors is rarely the case. CRL assumes causal relations between these latent factors making it more flexible and suitable for real world settings. In this work we give an overview over several approaches to disentangled representation learning and give a short introduction to variational auto-encoders and generative adversarial networks. We follow up by covering the current state-of-the-art CRL frameworks and finish with remarks regarding current weaknesses of CRL as well as potential research topics.

1 Introduction

In recent years a myriad of machine learning approaches were developed with the goal of solving all kinds of different problem settings, including, among others, tasks of classification [25], regression [8] and clustering [23]. All of

these different machine learning tasks have a heavy performance dependency on the choice of data (or feature) representation in common. Feature engineering offers a way to incorporate prior expert knowledge into data, leveraging human ingenuity at the cost of exceedingly labor intensive work. Representation learning is a broad term comprising techniques that learn data representations, from which it is easier to extract valuable information. In a sense, obtaining a more useful representation can be understood as gaining a better understanding of the underlying physical or logical mechanisms that produce the data. For example an image of an arbitrary object such as an apple. A human does not need to see the entirety of an image, as i.e. the outline of the apple will be sufficient for classification. Other meaningful characteristics could be e.g. an object's colour and its surface characteristics. Therefore, useful feature representations are generally of lower dimension than the original data but provide higher level information. Studies have shown that disentangling of the feature representations leads to better generalization and performance in neural networks [10, 28]. This aligns with the intuition that complex data is given birth by the rich interactions of comparably few explaining factors. At the same time the interpretability of the neural networks is improved, as changes in the output can often be traced back to a single or a small number of explaining factors. The challenge and topic of much research lies in developing methods that infer these disentangled representations. Disentangled representation learning (DRL) has many benefits but it assumes independent underlying factors. While this assumption holds or offers a sufficiently close approximation for many tasks, it does not for many more complex real world settings. Instead, the explanatory factors are often causally related. In economics, for example inflation, interest rates, employment levels, and consumer spending are fundamentally independent factors but can drastically influence each other. Another example are e.g. shadows in an image that depend on the positions of light sources as well as the shapes and positions of the illuminated objects. All of these causal relations are generally ignored when learning a typical disentangled representation. By enforcing these additional, causal restraints on the neural networks during training, further performance improvements can be achieved. Despite major efforts in DRL, as well as in research on causal discovery, very few studies have been done on causal representation learning (CRL). The scope of this survey includes a brief overview of

neural network-based DRL techniques, a survey on studies tackling CRL and finally an outlook over possible future research on the topic.

2 Disentangled Representation Learning

Representation learning and especially DRL has been a research topic of much interest over the last ten years. Bengio et al. [2] defined a disentangled representation as a number of distinct, independent, informative and generative factors, which are invariant to change in other generating factors. A more mathematical definition based on group theory was later proposed by Higgins et al. [11]. Several different frameworks exist for inferring disentangled representations. The majority of those are based on either the variational auto-encoder (VAE), proposed by Kingam and Welling [14] or the generative adversarial network (GAN), proposed by Goodfellow et al. [9]. In the following we will give a brief overview over relevant VAE and GAN-based models. A comprehensive overview of the topic of DRL was published by Wang et al. [28].

2.1 Variational Auto-Encoder

VAEs combine the structure of an auto-encoder [17] with variational inference. A stochastic encoder learns the parameters ϕ of the variational posterior distribution in order to map the data distribution $\mathbf{x}$ to the latent representation $\mathbf{z}$. The generative decoder on the other hand attempts to recreate the input given the latent representation by learning the parameters θ of the joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$, where $p_\theta(\mathbf{z})$ is the prior distribution over the latent space. The VAE attempts to maximize the log-likelihood of the data distribution $\log p_\theta(\mathbf{x})$ (2.1).

$$\log p_\theta(\mathbf{x}) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||(p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}_{\theta,\phi}(\mathbf{x}) \tag{2.1}$$

The by definition non negative Kulback-Leibler (KL) divergence D_{KL} pushes the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ to approximate the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$. $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ is the evidence lower bound (ELBO). In practice

the ELBO (2.2) is maximized in order to generate a tight bound on the log-likelihood $\log p_\theta(\mathbf{x})$.

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{x}|\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \tag{2.2}$$

The first term in equation 2.2 penalises the distance between the variational $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior distribution over the latent space $p_\theta(\mathbf{z})$. The second term on the other hand penalises the VAEs reconstruction error. Training such a VAE using stochastic gradient descent seems to be impossible at first due to the decoder's input $\mathbf{z}$ being sampled from the prior distribution over the latent space. The so called Reparameterization Trick by Kingma and Welling [14] bypasses this issue by multiplying the variance σ with a randomly sampled $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ instead of directly sampling $\mathbf{z}$.

While vanilla VAEs are generally able to infer disentangled representation, they rarely do without further constraints, especially when the data reaches a certain level of complexity. Higgins et al. [11] propose β-VAE, which introduces a penalty coefficient to the KL-divergence term in equation 2.2 which effectively encourages disentangled latent variables, but at the same time leads to a greater reconstruction error. Burgess et al. [4] approach the design of the loss function by simultaniously optimizing the mutual information between the input and task objective, as well as the mutual between the input and the latent space. In practice they subtract a linear parameter from the KL-divergence term in equation 2.2 and increases its value during network training. This leads to an improvement in the networks's reconstruction ability while still enforcing satisfactory disentanglement of the latent space. Kumar et al. [18] proposed DIP-VAE which, leverages an additional regularizer that penalizes the weighted distance between the marginal distribution of the variational posterior $q_\phi(\mathbf{z})$ and the latent prior $p_\theta(\mathbf{z})$. Kim et al. [13] introduced FactorVAE which utilizes the Total Correlation $D_{KL}(q_\phi(\mathbf{z})||\prod_j q_\phi(z_j))$, a measure of dimensional Independence, as an additional regularizer in the loss function. Chen et al. [5] propose β-TCVEA, in which they decompose the KL-divergence in equation 2.2 into three separately weighted terms as shown in equation 2.3.

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{x}|\mathbf{z}),p_\theta(\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \alpha I_q(\mathbf{z};\mathbf{x}) \\ - \beta D_{KL}(q_\phi(\mathbf{z})||\prod_j q_\phi(z_j)) - \gamma \sum_j D_{KL}(q_\phi(z_j)||p_\theta(z_j))\end{aligned} \tag{2.3}$$

α, β and γ are weights. The second term in equation 2.3 is the mutual information, the second term the Total Correlation and the third term a dimension wise KL-divergence.

2.2 Generative Adversarial Network

Another neural network architecture which researchers have adapted in order to infer disentangled representations is the GAN [9]. A GAN consists of two, with each other competing, neural networks. The generator network (G) creates data by sampling from a latent representation, while the discriminator network (D) attempts to differentiate real and generated data. As a result the two networks compete with each other in a zero-sum game. The training objective lies in the generator generating data which are indiscernible for the discriminator. Equation 2.4 shows a GAN's optimization objective.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim P_{Data}}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1 - D(G(\mathbf{z}))\right)\right] \tag{2.4}$$

P_{Data} is the real dataset and $p(\mathbf{z})$ is the prior distribution of the latent space. Chen et al. [6] proposed infoGAN, one of the first GAN-based approaches to DRL. The authors differentiate between a noise variable $\mathbf{z}$ and a structured latent variable $\mathbf{c}$. They then add a weighted regularizing term $-\lambda I(c; G(c,z))$ that encourages the mutual information between $\mathbf{c}$ and the generated data to stay large. Jeon et al. [12] introduce IB-GAN which compresses the inferred latent representation by limiting the mutual information, effectively achieving better disentangled representations. Lin et al. [19] propose a self supervised model InfoGAN-CR which uses a contrastive regularizer. The model is trained by evaluating generated images for which one generating latent factor is kept constant while all the others are randomly sampled. The contrastive regularizer then discerns the constant latent variable effectively encouraging the latent variables to be meaningful and distinct. Zhu et al. [34] proposes PS-SC GAN which utilizes spacial constrictions in the form of masks in order to localize the effect of latent variables in an image. Furthermore, in order for the disentangled latent space to encode simple and distinct variations in the data, they employ

perceptual simplicity by imposing small perturbations on single latent variables and identify them in the generated images. Finally Wei et al. [29] build a regularizer with a Jacobian of the generators output with respect to the latent input. The regularizer measures orthogonality of the Jacobian vectors, effectively rewarding independence of the latent dimensions.

3 Causal Representation Learning

Locatello et al. [21] argue that disentangled representations can only be inferred under inductive biases, both on the learning approach and on the data. They further query the mutual independence necessity of latent variables in DRL. CRL casts aside this independence assumption of the latent generating factors. Instead, the latent variables are dependent on each other, in accordance to an underlying causal mechanism. Shen et al. [26] proved that models with an independent latent prior distribution are not identifiable. CRL approaches condition their latent prior distributions by enforcing causal structure in different ways. Träuble et al. [27] further show that most DRL approaches are unable to disentangle latent factors if correlation exists in the data.

Yang et al. [30] were first to propose a representation learning framework which learns a structural causal model (SCM) [24] under light supervision. Their CausalVAE framework includes a SCM layer, which takes the independent exogenous factors $\boldsymbol{\epsilon}$ from the VAE decoder and transforms them into structured causal representations $\mathbf{z}$ which follow the structure of a directed acyclic graph (DAG). This DAG structure can be formulated as a strictly upper triangular adjacency matrix $\mathbf{A}$. The mathematical formulation of the SCM layer is shown in equation 3.1.

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{3.1}$$

$\mathbf{A}$ is the to be learned adjacency matrix and $\mathbf{I}$ the identity Matrix. In a next step, prior to being passed to the decoder, the causal representation $\mathbf{z}$ is passed to a Mask Layer where it reconstructs itself starting from the independent exogenous variables and following up with the on their parents dependent endogenous

variables according to the DAG structure. This process is described by equation 3.2.

$$z_i = g_i(\mathbf{A}_i \circ \mathbf{z}) + \epsilon_i \tag{3.2}$$

$\circ$ is the elementwise multiplication between the weight vectors $\mathbf{A}_i$ and $\mathbf{z}$ with $\mathbf{A} = [\mathbf{A}_1|...|\mathbf{A}_n]$ and g_i is a set of mildly nonlinear functions. The Mask Layer, besides enforcing the DAG characteristics on the causal representation $\mathbf{z}$, allows for interventions to be performed. In causality an intervention refers to the act of manipulating a variable in the system by external means. The outcomes of such interventions, besides being of interest themselves, can offer much insight on the underlying causal structure of a system. The weak supervision is realized through the labels u_i which hold the true causal concepts. They are fed to the encoder together with the data $\boldsymbol{\epsilon} = h(\mathbf{z}, \mathbf{u}) + \boldsymbol{\zeta}$ and further act as a constraint on the latent prior distribution $p_\theta(\mathbf{z}|\mathbf{u})$. Besides the ELBO objective (3.3) Yang et al. introduce three other regularization terms (3.4).

$$\begin{aligned} \textbf{ELBO} =& \mathbb{E}_{q\chi}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ & D_{KL}(q_\phi(\boldsymbol{\epsilon}|\mathbf{x},\mathbf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})) \\ & D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})||p_\theta(\mathbf{z}|\mathbf{u}))] \end{aligned} \tag{3.3}$$

$$\begin{aligned} \mathcal{L} = & - \textbf{ELBO} + \alpha \left(tr((\mathbf{I} + \frac{c}{m}\mathbf{A} \circ \mathbf{A})^n) - n \right) \\ & + \beta \mathbb{E}_{q\chi} \|\mathbf{u} - \sigma(\mathbf{A}^T\mathbf{u})\|_2^2 \\ & + \gamma \mathbb{E}_{\mathbf{z}\sim q_\phi} \sum_{i=1}^{n} \|z_i - g_i(\mathbf{A}_i \circ \mathbf{z})\|^2 \end{aligned} \tag{3.4}$$

α, β and γ are hyperparameters. The second term of equation 3.4 becomes zero iff the adjacency matrix $\mathbf{A}$ corresponds to a DAG. Zheng et al. [33] were the first to propose a continuous formulation of the DAG constraint for marticies. Yu et al. [31] then further developed the this formulation into the the form as shown in equation 3.4. The third and fourth term ease the training task of the two unknown variables $\mathbf{A}$ and $\mathbf{z}$ by encouraging $\mathbf{A}$ to correctly describe the causal relations of the labels $\mathbf{u}$, where σ is the logistic function and χ the data set, and $\mathbf{z}$ to be accurately reproduced by the Mask Layer. In addition to the CausalVAE

framework the authors provide a set of conditions under which their model is identifiable.

Komanduri et al. [16] propose the SCM-VAE framework. Contrary to the CausalVAE, SCM-VAE assumes that the underlying SCM is given and therefore does not need to be learnt during network training. Instead, it is utilized as a constrained on the latent prior distribution $p_\theta(\mathbf{z}|\mathbf{u})$. Similar to CausalVAE SCM-VAE incorporates a Mask Layer into the decoder, which allows for interventions to be performed on the SCM. Furthermore SCM-VAE is not limited by a linearity constraint on the SCM.

Recently, Komanduri et al. [15] improved on their SCM-VAE framework in the form of ICM-VAE. While they fundamentally assume the same given information, labels of the causal variables and the corresponding SCM, ICM-VAE improves on its predecessor by implementing independent causal mechanisms through learnable flow-based diffeomorphic functions. This implementation of structural causal flow allows for more complex models than the strictly additive noise model, which CausalVAE and SCM-VAE follow.

Similarly Fan et al. [7] propose another flow-based framework called Cauf-VAE, which does not require the underlying SCM as prior information, but infers it during training.

Brehmer and Haan et al. [3] propose an intervention-based framework for CRL. Instead of labels for the generating latent variables or the causal structure itself, they leverage datasets with paired samples from before and after random, unknown interventions. The authors define latent causal models (LCMs) as sets of an acyclic SCM $\mathcal{C}$, an observation space $\mathcal{X}$, a diffeomorphic decoder g, a set of interventions $\mathcal{I}$ on $\mathcal{C}$ and a probability measure $p_\mathcal{I}$ over $\mathcal{I}$. They then show, that under some mild constraints over the LCMs, two LCMs $\mathcal{M}$ and $\mathcal{M}'$ are equal up to a relabeling and an elementwise transformation of the causal variables if the LCMs entail equal weakly supervised distributions $p^\mathcal{X}_\mathcal{M}(\mathbf{x}, \tilde{\mathbf{x}}) = p^\mathcal{X}_{\mathcal{M}'}(\mathbf{x}, \tilde{\mathbf{x}})$ and vice versa. $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the observations before and after an intervention. Beyond a theoretical prove that a system's causal structure can be learnt under their weakly supervised setting, the authors introduce implicit latent causal models (ILCMs). Similar to the chicken-and-egg problem - which came first? - it is difficult to jointly learn the causal variables and the causal structure of a system

in an explicit fashion. The ILCM implicitly represents the causal structure through a neural solution function $s(\mathbf{e}) = \mathbf{z}$ which maps the vector of exogenous noise variables to the endogenous causal variables. The latent variables in an ILCM are therefore noise encodings defined by the inverse solution function. By learning the transformation $\tilde{\mathbf{z}}_i = \overline{s}(\tilde{\mathbf{e}}_i; \mathbf{e})$ in the ILCM, the solution function of a corresponding, unique explicit LCM (ELCM) can be recovered. the causal graph can be extracted from the ILCM after training by either using intervention-based causal discovery algorithms on the learned representations or by iterative topological ordering of the solution functions s_i.

An et al. [1] show, that in order to achieve causally sound generation through a decoder, a disentangled latent space is insufficient. Due to the entangled decoder, it is not given that the generated data indeed follows the inferred causal structure of the latent space. For instance, a do-operation on a single variable in latent space might affect that variable's parents, as the natural decoder has no regularization preventing this from happening. Their framework CDG-VAE allows for correct causal generation under the rather strict assumption of comprehensive knowledge over the causal information for supervision.

Lippe et al. [20] tackle the problem of CRL from temporal sequences. Previous works have made the naive assumption that intervening on a variable only has an effect in later time steps. This is often inaccurate, as there are often causal effects in real world settings which act faster than the modeled time steps. The authors show that by using interventions, effectively removing instantaneous effects of parent variables, they can recover the minimal causal variables and identify the caual graph under mild assumptions.

Louizos et al. [22] and Zhang et al. [32] provide representation learning frameworks for treatment effect estimation. Treatment effect estimation tries to predict the effect of a certain treatment under hidden confounding. Both works assume a surrogate rich setting, ample indirect information about the hidden confounding variable. While they are technically not learning a causal representation, their models CEVAE [22] and TEDVAE [32] hold the causal structure of the setting implicitly, and learn the latent hidden confounder during training. TED-VAE improves upon CEVAE by further distinguishing between multiple latent variables depending on their effect on either/and treatment and outcome.

While the field of CRL is mainly dominated by VAE-based approaches Shen et al. [26] introduced DEAR, a GAN-based CRL framework. They employ labels of the causal variables for supervision and assume a super-graph of the underlying causal graph is known. The DEAR's loss function is given in equation 3.5.

$$\min_{E,G,F} L_{E,G,F} := D_{KL}(q_E(\mathbf{x},\mathbf{z}), p_{G,F}(\mathbf{x},\mathbf{z})) + \lambda \mathbb{E}_{\mathbf{x},\mathbf{y}}[l_s(E;\mathbf{x},\mathbf{y})] \tag{3.5}$$

E references the encoder network, G the generative network and F the causal prior. The encoded joint distribution factorizes as $q_E(\mathbf{x},\mathbf{z}) = q_{\mathbf{x}}(\mathbf{x})q_E(\mathbf{z}|\mathbf{x})$ and the generated joint distribution as $p_{G,F}(\mathbf{x},\mathbf{z}) = p_F(\mathbf{z})p_G(\mathbf{x}|\mathbf{z})$, where

$$p_F(\mathbf{z}) = f((\mathbf{I} - \mathbf{A}^T)^{-1}h(\epsilon))$$

encodes the generally nonlinear SCM. f and h are generally nonlinear element-wise transformations, $\mathbf{A}$ is the weighted adjacency matrix corresponding to the causal graph and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. The second term of equation 3.5 is a supervised regularizer. λ is a hyperparameter, $\mathbf{y}$ are the supervision labels and l_s a function which penalizes deviation from the labels. In the case of binary labels the authors chose a cross entropy loss. Like many of the other CRL frameworks, DEAR allows for interventions to be performed on the generating latent variables.

4 Possible Research Ideas and Final Remarks

Even though CRL is still a rather young research topic it has obtained much interest from the scientific community. Compared to general DRL, CRL is much more applicable to real world problems, as causal relations between generating latent factors are much more frequent than the contrary. On the other hand, current CRL frameworks are limited by their need for supervision in order to guarantee identifiability. Labels for all latent causal variables or for example the complete causal graph are rather strict requirements which cannot always be fulfilled in real world settings. Similarly, datasets of paired samples from before and after interventions might be feasible for some of the generating variables but rarely for all. Therefore, a framework which allows for not only a single type

of supervision but a mixture of different types of supervision would make CRL frameworks much more applicable. Another scenario which, to the best of our knowledge, has not been covered yet, is the case of ample supervision on all but one generating variable. Current approaches require some form of supervision for all generating variables. In real world problems, such as manufacturing processes, this cannot be guaranteed as there might be further, hidden variables which have not been considered. In future research we plan to investigate CRL-based methods for inferring quantitative or even qualitative information on such hidden variables. Attaining knowledge about the existence or even better, about the placement of such an additional hidden variable within the causal graph would be very valuable. Once discovered, additional measurements or even an instrumentation could be done in order to obtain additional supervision labels.

References

[1] SeungHwan An, Kyungwoo Song, and Jong-June Jeon. "Causally Disentangled Generative Variational AutoEncoder". In: *ArXiv* abs/2302.11737 (2023). URL: https://api.semanticscholar.org/CorpusID:257102874.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[3] Johann Brehmer et al. "Weakly supervised causal representation learning". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 38319–38331. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/fa567e2b2c870f8f09a87b6e73370869-Paper-Conference.pdf.

[4] Christopher P Burgess et al. "Understanding disentangling in β-VAE". In: *arXiv preprint arXiv:1804.03599* (2018).

[5] Ricky TQ Chen et al. “Isolating sources of disentanglement in variational autoencoders”. In: *Advances in neural information processing systems* 31 (2018).

[6] Xi Chen et al. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets”. In: *Advances in neural information processing systems* 29 (2016).

[7] Di Fan, Yannian Hou, and Chuanhou Gao. “CF-VAE: Causal Disentangled Representation Learning with VAE and Causal Flows”. In: *arXiv preprint arXiv:2304.09010* (2023).

[8] Manuel Fernández-Delgado et al. “An extensive experimental survey of regression methods”. In: *Neural Networks* 111 (2019), pp. 11–34.

[9] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[10] Irina Higgins et al. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: *International conference on learning representations*. 2016.

[11] Irina Higgins et al. “Towards a Definition of Disentangled Representations”. In: *CoRR* abs/1812.02230 (2018). arXiv: 1812.02230. URL: http://arxiv.org/abs/1812.02230.

[12] Insu Jeon et al. “Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 9. 2021, pp. 7926–7934.

[13] Hyunjik Kim and Andriy Mnih. “Disentangling by factorising”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2649–2658.

[14] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[15] Aneesh Komanduri et al. “Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms”. In: *arXiv preprint arXiv:2306.01213* (2023).

[16] Aneesh Komanduri et al. “SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge”. In: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. 2022, pp. 1014–1023.

[17] Mark A Kramer. “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE journal* 37.2 (1991), pp. 233–243.

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. “Variational inference of disentangled latent concepts from unlabeled observations”. In: *arXiv preprint arXiv:1711.00848* (2017).

[19] Zinan Lin et al. “Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans”. In: *international conference on machine learning*. PMLR. 2020, pp. 6127–6139.

[20] Phillip Lippe et al. “Causal representation learning for instantaneous and temporal effects in interactive systems”. In: *The Eleventh International Conference on Learning Representations*. 2022.

[21] Francesco Locatello et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[22] Christos Louizos et al. “Causal effect inference with deep latent-variable models”. In: *Advances in neural information processing systems* 30 (2017).

[23] Erxue Min et al. “A survey of clustering with deep learning: From the perspective of network architecture”. In: *IEEE Access* 6 (2018), pp. 39501–39514.

[24] Judea Pearl. *Causality*. Cambridge university press, 2009.

[25] Waseem Rawat and Zenghui Wang. “Deep convolutional neural networks for image classification: A comprehensive review”. In: *Neural computation* 29.9 (2017), pp. 2352–2449.

[26] Xinwei Shen et al. “Weakly supervised disentangled generative causal representation learning”. In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 10994–11048.

[27] Frederik Träuble et al. “On disentangled representations learned from correlated data”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10401–10412.

[28] Xin Wang et al. “Disentangled representation learning”. In: *arXiv preprint arXiv:2211.11695* (2022).

[29] Yuxiang Wei et al. “Orthogonal jacobian regularization for unsupervised disentanglement in image generation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6721–6730.

[30] Mengyue Yang et al. “Causalvae: Disentangled representation learning via neural structural causal models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9593–9602.

[31] Yue Yu et al. “DAG-GNN: DAG structure learning with graph neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7154–7163.

[32] Weijia Zhang, Lin Liu, and Jiuyong Li. “Treatment effect estimation with disentangled latent factors”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 10923–10930.

[33] Xun Zheng et al. “Dags with no tears: Continuous optimization for structure learning”. In: *Advances in neural information processing systems* 31 (2018).

[34] Xinqi Zhu, Chang Xu, and Dacheng Tao. “Where and what? examining interpretable disentangled representations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5861–5870.

Bayesian Optimization of Immature Multi-Stage Processes

Saksham Kiroriwal

Kognitive Industrielle Systeme (KIS)
Fraunhofer IOSB, Germany
saksham.kiroriwal@iosb.fraunhofer.de

Abstract

The optimization of chained production processes is a fundamental and continually evolving challenge in modern manufacturing. This paper discusses the application of Bayesian optimization to the optimization of production processes, where the output of one stage directly influences the parameters of subsequent stages. Current research work explores concurrent exploration and optimization, cost considerations, and the integration of expert knowledge and multi-stage process models. This paper not only provides insights into the mathematical modeling of the problem, but also provides insights into the recent advancements for the same. This research opens the door to a range of applications and further exploration in the field of production process optimization.

1 Introduction

In the field of modern manufacturing, the optimization of production processes is a fundamental and continually evolving challenge. These processes have become increasingly complex with time. The advancement of manufacturing no longer depends solely on physical hardware, but also on the intricate interaction of software, control systems, and process parameters [4]. In this context, we

are concerned with the optimization of chained production processes, which resemble interconnected function networks [3]. These processes introduce a unique challenge, as the output of one function directly influences the input of the next. An example of a two-stage production process has been in Fig. 1.1

The process consists of subprocesses $P^{(n)}$ and comprises of the controller $\pi^(n)$ and the process state $\mathbf{s}^{(n)}$, where $n \in \{1, \ldots, N\}$. Each process subprocess takes an input vector $\mathbf{x} \in \mathbb{R}^{d_x^{(n)}}$ and produces an output(s) $\mathbf{y} \in \mathbb{R}^{d_y^{(n)}}$. The software update version of a process is indicated by z, while the iteration / number of runs of the processes is indicated by s.

An example process could be a stamp-forming for carbon fiber sheet process in which the subprocesses $P^{(1)}, P^{(2)}, P^{(3)}$ could represent laying of carbon fiber tapes, heating and forming. Each of the subprocesses has a different set of input parameters $\mathbf{x}^{(n)}$. $\mathbf{x}^{(1)}$ would include the tape-laying pattern, the thickness and width of the tape, $\mathbf{x}^{(2)}$ would include the heating temperature and time ,and $\mathbf{x}^{(3)}$ would include the stamp force and the shape of the stamp.

The state-of-the-art in manufacturing is heavily dependent on elements beyond physical machinery. The maturity of a manufacturing process depends on the refinement of automation software, feedback-control systems, and fixed process parameters. But a process can become immature or sub-optimal when there are any changes to the controller, raw material or a different final product is required. The focal point is the rapid maturation of production processes achieved through the optimization of the fixed process parameters $\mathbf{x} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ within individual production runs. These parameters, represented as finite-dimensional vectors, play a pivotal role in ensuring efficient and high-quality production [4]. For experiments / research, the stability of the hardware, automation software, and control systems is assumed. We also presume that the process is capable of producing products, albeit with low maturity.

Optimizing these fixed parameters is no small feat as it requires minimizing resource usage in terms of both time and material cost. [10, 15] Traditional optimization methods often rely on heuristics or theoretical approaches, which, while effective in many cases, can lead to local minima. The introduction of large-scale machine learning methods has opened new avenues for addressing these challenges. Black-Box Optimization, a class of problems where the objective

function is initially unknown, has received significant attention[4, 10]. These methods make minimal assumptions about the objective function, including nature-inspired heuristics such as differential evolution [22] and methods that provide good convergence to global optimal [12].

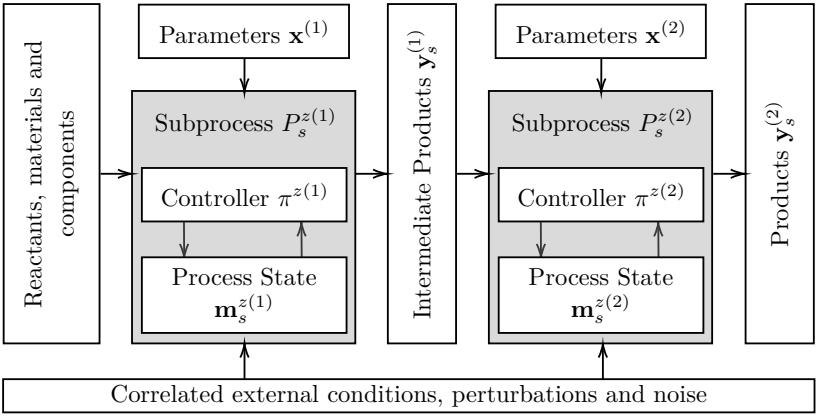


Figure 1.1: Conceptual view of a two-stage production process

To tackle this intricate optimization problem, we turn to Bayesian optimization (BO), a powerful gradient-free approach for enhancing the efficiency and quality of production systems. Bayesian optimization has demonstrated its versatility by offering solutions not only for single and multi-objective optimization [16], but also for policy search in reinforcement learning [25], the optimization of complex systems modeled through simulation, and the calibration of time-consuming physics-based models (Snoek et al. [20, 21]; Frazier et al. [9]). These capabilities make it a promising tool to address the intricate challenges of chained production processes. Recent developments in Bayesian optimization, particularly in scenarios where inequality constraints need to be applied, have enabled the seamless integration of constraints into the optimization process. Furthermore, the intricate issue of multi-stage robust optimization, which allows partial observation of intermediate results, requires more intricate modeling and optimization policy.

The primary objective of this paper is to discuss the state of the art application of Bayesian optimization to chained production processes and possible research areas. The focus is on the maturation of multi-stage production processes, where the output of one stage significantly influences the parameters of subsequent stages. We intend to develop methods that facilitate concurrent exploration and optimization, taking into account cost considerations and the potential for scrap production. The motivation for this study comes from the composite stamp-forming process, used as a validation case. Nevertheless, the research aims to develop a generalized methodology for application to various production processes and analogous problems beyond this project's scope.

Section 2.1 discussed the background on Gaussian Process(GP) and Bayesian Optimization (BO). Section 3 explains the mathematical formulation of the problem statement, key challenges of multi-stage process optimization and some related work which will also become the starting point for further research.

2 Bayesian Optimization

Bayesian optimization, grounded in Bayesian decision theory, represents a powerful methodology for optimizing black-box objective functions with inherent evaluation challenges, whether due to physical limitations, computational complexity, or financial constraints. Bayesian Optimization uses Gaussian processes (GPs) to sequentially optimize expensive objective functions, offering a systematic and efficient approach to tasks such as hyperparameter tuning, experimental design, and robotics. By iteratively selecting points to evaluate and updating the GP model, Bayesian Optimization strikes a balance between exploration and exploitation, making it a valuable tool for tackling optimization challenges in various domains.

2.1 Gaussian Processes (GP)

GPs are powerful probabilistic models that capture complex relationships in data. They provide not only predictions, but also uncertainty estimates, making them ideal for modeling and optimizing costly real-world processes. Linear

regression follows the "weight space view", where it is assumed that each data point, y_s, is generated from the corresponding $f(\mathbf{x}_s)$ and some independent Gaussian noise [7, 18], represented as:

$$y_s = f(\mathbf{x}_s) + \epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, \sigma_\epsilon^2 I). \tag{2.1}$$

The latent function $f : \mathcal{X} \to \mathcal{Y}$ is drawn from a Gaussian process prior, represented as $f \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}^{'}))$. [19] The prior relies solely on the covariance function, denoted as k, operating on the input vector $\mathbf{x}$. In pursuit of a flexible model, we refrain from making specific assumptions about the form of the generative mapping f. Instead, the choice of covariance function defines the essential properties of this mapping. For instance, consider the exponentiated quadratic covariance function for two vector inputs $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$k(\mathbf{x}_i, \mathbf{x}_j) = k_{\mathbf{x}_i, \mathbf{x}_j} = (\sigma_{se})^2 \exp\left(-\frac{|\ \mathbf{x}_i - \mathbf{x}_j\ |^2}{2\lambda^2}\right), \tag{2.2}$$

which imparts infinite smoothness to latent functions. [18] The hyperparameters of the covariance function, such as (σ_{se}, λ), are collectively represented by $\boldsymbol{\theta}$. Let data $\mathcal{D}$ consist of input matrix $\mathbf{X} \in \mathbb{R}^{S \times R}$, output vector $\mathbf{y} \in \mathbb{R}^S$ and test input $\mathbf{x}_* \in \mathbb{R}^R$. The joint distribution of the latent function at the test point, f_*, and observed outputs can be expressed under the prior as

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} = \mathcal{N} \sim \left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_\epsilon^2 \mathbf{I} & \mathbf{k}^T \\ \mathbf{k} & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right) \tag{2.3}$$

Using a conditional distribution, the prediction f_* at the test input can be calculated using Bayesian inference on the GP. The Gaussian nature allows the

posterior predictive to be computed analytically as [18, 19]

$$p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \mathcal{N} \sim (\mu, \Sigma)\,, \tag{2.4a}$$

$$\mu = \mathbf{k}^T \left(\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right)^{-1} \mathbf{y}, \tag{2.4b}$$

$$\Sigma = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^T \left(\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right)^{-1} \mathbf{k}. \tag{2.4c}$$

$$\text{where } \mathbf{k} = k(\mathbf{x}_*, \mathbf{X}) = \begin{bmatrix} k(\mathbf{x}_*, \mathbf{x}_1) \\ \vdots \\ k(\mathbf{x}_*, \mathbf{x}_S) \end{bmatrix} \text{ and,} \tag{2.4d}$$

$$\mathbf{K} = k(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_S) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_S, \mathbf{x}_1) & \cdots & k(\mathbf{x}_S, \mathbf{x}_S) \end{bmatrix} \tag{2.4e}$$

The distribution of the output at the test location with additive Gaussian noise is expressed as

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) \sim \mathcal{N}\left(\mu, \Sigma + \sigma_\epsilon^2\right). \tag{2.5}$$

This process allows for valuable insights into the relationships within the data, enabling probabilistic inference for the output.

2.2 Bayesian Decision Theory

Bayesian optimization effectively addresses a sequential decision problem characterized by a finite horizon T. [10] While conventional Bayesian optimization is well suited for optimizing functions within continuous domains, typically characterized by fewer than 20 dimensions, recent advancements, as elucidated by Eriksson et al. [8], have expanded its applicability even to higher-dimensional spaces, such as those with 40 dimensions.

The fundamental concept of Bayesian optimization revolves around the development of a surrogate model, which serves as a representation of the objective function, coupled with the quantification of the uncertainty associated with this function. [26] As discussed in Section 2.1, the Gaussian process prior (GP), the most widely used surrogate, provides a probabilistic model to characterize the relationships between input-output pairs. Given the observed data $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$,

a distribution $p(y_*|\mathbf{x}_*, \mathcal{D})$ for the output y_* at an unobserved point $\mathbf{x}_*$ can be estimated.

Central to Bayesian decision theory lies the concept of a utility function, symbolized as $u(\mathcal{D})$, which quantifies the usefulness of the observed data $\mathcal{D}$. In the domain of Bayesian optimization, this utility function effectively explains the information obtained from each experimental observation. $u(y_*|\mathbf{x}_*, \mathcal{D})$ represents the marginal gain in utility after observing a new data point y_* at input $\mathbf{x}_*$, given the observed data $\mathcal{D}$. This gain captures the additional value obtained by acquiring this new information. It is defined as [10, 11]

$$u(y_*|\mathbf{x}_*, \mathcal{D}) = u(\mathcal{D} \cup \{(\mathbf{x}_*, y_*)\}) - u(\mathcal{D}). \tag{2.6}$$

For the optimization task, when there are t steps remaining, the expected utility $Q_t(\mathbf{x}_*|\mathcal{D})$ quantifies the expected utility derived from the experiment at new location $\mathbf{x}_*$, given the observed data $\mathcal{D}$ and the predefined number of remaining steps, t. It is expressed through the Bellman equation as follows:

$$Q_t(\mathbf{x}_*|\mathcal{D}) = \mathbb{E}[u(y_*|\mathbf{x}_*, \mathcal{D})] + \mathbb{E}[\max_{x'} Q_{t-1}(x'|\mathcal{D} \cup \{\mathbf{x}_*, y_*\})]. \tag{2.7}$$

Here, the expectation is computed with respect to $p(y_*|\mathbf{x}_*, \mathcal{D})$. The Bellman equation iteratively evaluates the expected utility linked to potential actions, comprising immediate and expected future utility gain. The optimal input

$$\mathbf{x}_* = \arg\max_{\mathbf{x} \in \mathcal{X}} Q_t(\mathbf{x}|\mathcal{D}), \tag{2.8}$$

corresponds to the input that maximizes the expected utility increase for t remaining steps. [10, 11]

The search for an optimal policy in Bayesian optimization is based on maximizing the expected utility across a predetermined horizon. However, finding the globally optimal policy can be computationally intractable. In practice, the horizon is kept finite ($T = 1$ or 2) and the policy selects the action that promises the highest immediate expected utility. When $T = 1$, it is known as a one-step look-ahead policy. [10, 11]

The selection of the utility function represents a pivotal decision with implications for the exploration-exploitation trade-off within the domain of Bayesian

optimization. A commonly used utility function is the "expected improvement (EI)," where, given an initial set of observations $\mathcal{D}_o$ with the best-observed output y^* and a new input $\mathbf{x}'$, the utility function, or simple reward, is defined as the improvement over y^*, expressed as

$$EI(\mathbf{x}) = \max(\mu_{\mathcal{D}}(\mathbf{x}') - y^*, 0), \tag{2.9}$$

where $\mu_{\mathcal{D}}(\mathbf{x}')$ denotes the posterior mean. [10, 11]

In Bayesian optimization, the expected utility soon becomes intractable when noisy observations and noisy inputs are present. As shown in Fig. 1.1, the output of a stage $n \in \{1, \ldots, N-1\}$ becomes the input for the next stage, due to which uncertainty in the input also needs to be modelled in the characterizing function for the nodes. This makes the GP posterior intractable, and hence also the expected utility intractable. To address this issue, approximation methods like sample average approximation in combination with reparameterization trick need to be used as mentioned by Astudillo et al. [3] and Kusakawa et al. [14]. The partial observability of Directed Acyclic Graph (DAG) also provides an interesting insight into the inference from the cascaded / network GP as marginalization over all the latent layers is not required, like in the case of deep GP [6]. The layers in the discussed DAG are conditionally independent as the intermediate outputs can be observed. [3] Section 3 discusses some implemented ideas that will lay down the foundation for future work.

3 Multi-Stage Bayesian Optimization

In this section, we formalize the problem of optimizing multi-stage production processes represented as a function network and also discuss some work that has been done previously. The multi-stage process is conceptualized as a Directed Acyclic Graph (DAG) as shown in Fig. 3.1. The function network consists of N nodes or stages, where each node $n \in \{1, \ldots, N\}$ represents a subprocess. [3, 14]

3.1 Problem Statement

For each node n, let $\mathcal{J}^{(n)}$ denote the set of parent nodes of node n. We assume that the nodes are ordered in such a way that $j < n$ for all $j \in \mathcal{J}^{(n)}$. Furthermore, each node accepts a set of controllable inputs $\mathbf{x}^{(n)} \in \mathbb{R}^{d_x^{(n)}}$ and outputs from the parent nodes $\mathcal{Y}^{(n)}_{\mathcal{J}^{(n)}}$ and produces an output $y^{(n)} \in \mathbb{R}$. Here, $d_x^{(n)}$ denotes the number of controllable input parameters for node n.

The behavior of each node is modeled by a function $f^{(n)} : \mathcal{X}^{(n)} \times \mathcal{Y}^{(n)}_{\mathcal{J}^{(n)}} \to \mathcal{Y}^{(n)}$. This relationship can be expressed as:

$$y^{(n)} = f^{(n)} \left(\mathbf{x}^{(n)}, y^{(n)}_{\mathcal{J}^{(n)}} \right)$$

with the initial condition $\mathcal{Y}^{(0)}_{\mathcal{J}^{(0)}} = \{\emptyset\}$ and $y^{(0)}_{\mathcal{J}^{(0)}} = \emptyset$ for the first node.

This representation signifies the interdependence between stages, where the output of one or more stage(s) serves as the input for the subsequent stage(s). The objective is to find the decision vector $\mathbf{x}$ such that

$$\arg \max_{\mathbf{x} \in \mathcal{X}^{(N)}} f^{(N)} \left(\mathbf{x}, y^{(N)}_{\mathcal{J}^{(N)}} \right),$$

thus optimizing the multi-stage process.

The optimization of the multistage production process presents several challenges. Notably, the high dimensionality of the parameter space and the intricacies of sequential decision-making require innovative approaches. Additionally, the time-consuming nature of evaluating functions within the network necessitates efficient optimization techniques. The network structure of the multi-stage process demands consideration of the uncertainty in the output from the previous stage. It is not solely a question of optimizing the parameters $\mathbf{x}^{(n)}$ at each stage but also a question of determining the optimal node for optimization. The choice of which stage to optimize can significantly impact the overall process performance.

Gaussian processes (GPs) emerge as a valuable probabilistic model for Bayesian optimization in this context. [3, 14] GPs provide a flexible framework for modeling the underlying functions $f^{(n)}$. They not only offer a surrogate model for

function values, but also quantify the uncertainty associated with these predictions. The probabilistic nature of GPs is particularly advantageous when dealing with uncertainty in complex multi-stage processes. [26] Due to this reason, GPs have been used to model the function of nodes/stages in the network.

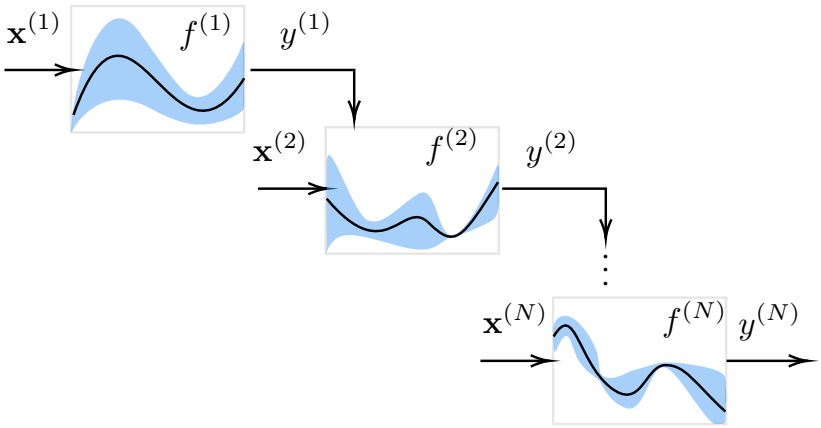


Figure 3.1: DAG representation of multi-stage process

The problem at hand goes beyond conventional optimization tasks due to its multi-stage nature, function interdependencies, and the need for efficient exploration and optimization techniques in high dimensions. In the subsequent sections, related work in the application of Bayesian optimization, utilizing Gaussian processes, to address these challenges has been discussed.

3.2 Related Work

The work of Zhilinskas et al. [28], Močkus et al. [17] and Jones et al. [12] laid the foundation of Bayesian optimization. Recent advances in Bayesian optimization and computational advancements have led to increased application in machine learning hyperparameter tuning and black-box optimization. The works by

Snoek et al. [20, 21], Balandat et al. [5], Sui et al. [23], Wilson et al. [27] and Jiang et al. [11] provide an insight to the state of the art developments.

The work of Astudillo et al. [3] and Kusakawa et al. [14] focuses on the application of parameter space exploration using Bayesian optimization for function networks shown in Fig. 3.1. Instead of modeling a complete black box, they leverage the DAG structure of the production process and incorporate a grey box methodology for process optimization. However, it should be noted that there exists a subtle difference between the two methodologies for Bayesian optimization for multi-stage process. The work from the two has been discussed in this section.

Astudillo et al. [3] describes the optimization in three steps which includes the calculation of posterior for the final stage, expected improvement of the function network and the maximization of the function network. Rather than modeling the network directly, the node functions, $f^{(1)}, \ldots, f^{(N)}$, are drawn from independent GP prior distributions. Let prior mean and covariance functions for $f^{(n)}$ be defined as $\mu_0^{(n)}, \Sigma_0^{(n)}$. As in this section, due to the interdependence and observability of node outputs, if $f^{(N)}$ is calculated S times, $f^{(1)}, \ldots, f^{(N-1)}$ are also observed S times. Thus, if f^N is queried S times, posterior distributions from independent GPs for $f^{(1)}, \ldots, f^{(N-1)}$ with mean and covriance functions $\mu_S^{(n)}$ and $\Sigma_S^{(n)}$ can be obtained using GP regression [18] equations. [3]

The output $y_s^{(n)}$ for a node n is considered a sample drawn from the posterior predictive distribution of the GP for $f^{(n)}$. The acquisition function "Expected Improvement" for the network is defined as the improvement from current input $\mathbf{x} \in \{\mathcal{X}^{(1)} \cup \ldots \mathcal{X}^{(N)}\}$ over the so far best observed output from the last node

$$\text{EI-FN}_S(\mathbf{x}) = \mathbb{E}_S \left[\max \left(\{y^{(N)} - y_S^{*(N)}\}, 0 \right) \right], \tag{3.1}$$

where $y_S^{*(N)} = \max_{s \in \{1,\ldots,S\}} y_s^{(N)}$. However due to network of functions, neither posterior distribution of $f^{(N)}$ nor acquisition function can be calculated in closed form. This is solved using reparameterization trick [27, 13] and sample average approximation as shown in [5, 3]. Samples from the obtained posteriors can be calculated to approximate the distribution for a particular function. The

intractable EI-FN$_S(\mathbf{x})$ is approximated using Monte-Carlo (MC) approach [3]

$$\text{EI-FN}_S(\mathbf{x}) \approx \frac{1}{Q}\sum_{q=1}^{Q}\left[\max\left(\{y^{(N)} - y_S^{*(N)}\}, 0\right)\right]. \tag{3.2}$$

One crucial observation is that the input for all the stages is chosen at the very start of the multi-stage process, and improvement in a subprocess cannot be made while the process chain is underway. It is also‘ the key difference between the work by Astudillo et al. and Kusakawa et al.

Kusakawa et al. [14] define the acquisition function in such a way that the input to a particular node/stage can be changed without rerunning the previous stages. The utility function for the last node N is defined in similar way as the EI-FN$(\mathbf{x})$ as

$$U^{(N)}\left(\mathbf{x}^{(N)}, y^{(N)}_{\mathcal{J}^{(N)}}\right) = \mathbb{E}_{f^{(N)}}\left[\max\left(\{y^{(N)} - y_S^{*(N)}\}, 0\right)\right]. \tag{3.3}$$

On the other hand, the utility function for intermediate nodes $n < N$ is defined as

$$U^{(n)}\left(\mathbf{x}^{(n)}, y^{(n)}_{\mathcal{J}^{(n)}}\right) = \mathbb{E}_{f^{(n)}}\left[\max_{\mathbf{x}\in\mathcal{X}^{(n+1)}} U^{(n+1)}\left(\mathbf{x}, y^{(n+1)}_{\mathcal{J}^{(n+1)}}\right)\right]. \tag{3.4}$$

Writing the recursive maximization and expectation, the utility for intermediate nodes can be approximate as

$$\begin{aligned} U^{(n)}\left(\mathbf{x}^{(n)}, y^{(n)}_{\mathcal{J}^{(n)}}\right) = \mathbb{E}_{f^{(n)}}\Bigg[&\max_{\mathbf{x}\in\mathcal{X}^{(n+1)}} \dots \\ &\mathbb{E}_{f^{(N-1)}}\left[\max_{\mathbf{x}\in\mathcal{X}^{(N)}} U^{(N)}\left(\mathbf{x}, y^{(N)}_{\mathcal{J}^{(N)}}\right)\right]\Bigg]. \end{aligned} \tag{3.5}$$

Equation 3.5 takes form of the famous Bellman equation. The lower bound of the recursive Bellman equation is determined using insights from the work of Jiang et al. [11]. Balandat et al. [5] provide the MC approximation methods for intractable distribution. The approximated expected utility can be approximated

as

$$U^{(n)}\left(\mathbf{x}^{(n)}, y^{(n)}_{\mathcal{J}^{(n)}}\right) \geq \max_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathcal{F}}\left[U^{(N)}\left(\mathbf{x}, y^{(N)}_{\mathcal{J}^{(N)}}\right)\right] \tag{3.6a}$$

$$\approx \max_{\mathbf{x}\in\mathcal{X}} \frac{1}{Q}\sum_{q=1}^{Q} U^{(N)}\left(\mathbf{x}, \hat{y}^{(N)}_{\mathcal{J}^{(N)}}\right) \tag{3.6b}$$

where $\mathcal{X} = \{\mathcal{X}^{(n+1)} \cup \ldots \mathcal{X}^{(N)}\}$ and $\mathcal{F} = f^{(n)}, \ldots, f^{(N-1)}$. It can be noticed that the utility of node n is approximated using the utility of the last node N and in turn requires the intermediate and final outputs $y^{(n+1)}_{\mathcal{J}^{(n+1)}}, \ldots, y^{(N)}_{\mathcal{J}^{(N)}}$. Since these are not available at node n, they need to be to approximated using GP network. The calculated intermediate outputs $\hat{y}^{(n+1)}_{\mathcal{J}^{(n+1)}}, \ldots, \hat{y}^{(N)}_{\mathcal{J}^{(N)}}$ are found in same way as explained in earlier in this section and sampled from the posterior predictives of the GPs. As discussed in [11], GPs are an approximation to the original latent functions and looking too far into the future utility can also lead to subpar results. So a balance between the foresightedness and approximation needs to be considered when using the methodology by Kusakawa et al.

The problem formulation by the two discussed work and the supporting advancements can serve as a foundation for future work in the analysis of optimization methods for function networks. The work not only provides a mathematical description of the problem but also the state of the art work in the similar field.

4 Research Gap and Future Work

The work by Astudillo et al. [3] and Kusakawa et al. [14] provides insights into how expected improvement can be formulated for a DAG. But as the process becomes larger, the input dimension becomes too large to be for the inner optimization loop of maximization of acquisition function. This requires methods for dimensionality reduction and efficient sampling. Also, the acquisition functions become intractable due to the cascaded GPs and requires some approximations for the same. Study can be done to investigate the effect of different approximation methods like sample average approximation with different sampling methods. Also, the mentioned studies still work in an environment where the decision making is non-myopic with respect to the multi-stage process but

myopic to the subprocess. Approximate dynamic programming methods could be investigated to develop methodologies that can be non-myopic with respect to the subprocess too.

Balandat et al. [5] and Jiang et al. [11] provide methods that are applicable for single-stage process and are non-myopic. They try to find a lower bound for non-myopic expected marginal utility gain that tighter. Balandat et al. [5] provide one-shot method for approximation of non-myopic marginal utility gain for single-stage optimization and nicely shows how different algorithms in past approximate the lower bounds and how good the approximation is. Using this, investigation can be done to find out lower bound for non-myopic utility gain for multi-stage process. This could provide open the gates to one-shot Bayesian optimization for multi-stage process.

Looking from a different perspective of causal analysis, some recent work by Aglietti et al. [1, 2] and Sussex et al. [24] introduces a new take on Bayesian optimization by incorporation causal knownledge into the acquisition function and reducing the dimensionality of the search space. They use the concept of possibly optimal minimal intervention set and intervention-observation tradeoff for reducing the search space while incorporating causal intervention knowledge about the system. Investigation can be done regarding the comparison of different causal and non-causal methodologies.

In conclusion, the optimization of chained production processes presents a unique and complex challenge in contemporary manufacturing. Bayesian optimization emerges as a powerful tool for addressing this challenge and advancing the maturation of multi-stage production processes. The work discussed in Section 3 can serve as a building block and foundation for further research. Work can be done to formulate methods that approximate the tighter lower bound of the Bellman equation or develop better acquisition functions which can exploit the graph of the available process.

References

[1] Virginia Aglietti et al. “Causal bayesian optimization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3155–3164.

[2] Virginia Aglietti et al. “Dynamic causal Bayesian optimization”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10549–10560.

[3] Raul Astudillo and Peter Frazier. “Bayesian optimization of function networks”. In: *Advances in neural information processing systems* 34 (2021), pp. 14463–14475.

[4] Charles Audet and Warren Hare. “Derivative-free and blackbox optimization”. In: (2017).

[5] Maximilian Balandat et al. “BoTorch: A framework for efficient Monte-Carlo Bayesian optimization”. In: *Advances in neural information processing systems* 33 (2020), pp. 21524–21538.

[6] Andreas Damianou and Neil D Lawrence. “Deep gaussian processes”. In: *Artificial intelligence and statistics*. PMLR. 2013, pp. 207–215.

[7] Issam El Naqa and Martin J Murphy. *What is machine learning?* Springer, 2015.

[8] David Eriksson and Martin Jankowiak. “High-dimensional Bayesian optimization with sparse axis-aligned subspaces”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 493–503.

[9] Peter I Frazier. “A tutorial on Bayesian optimization”. In: *arXiv preprint arXiv:1807.02811* (2018).

[10] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

[11] Shali Jiang et al. “BINOCULARS for efficient, nonmyopic sequential experimental design”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4794–4803.

[12] Donald R Jones, Matthias Schonlau, and William J Welch. “Efficient global optimization of expensive black-box functions”. In: *Journal of Global Optimization* 13.4 (1998), pp. 455–492.

[13] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[14] Shunya Kusakawa et al. “Bayesian Optimization for Cascade-Type Multistage Processes”. In: *Neural Computation* 34.12 (Nov. 2022), pp. 2408–2431. ISSN: 0899-7667.

[15] Remi Lam, Karen Willcox, and David H Wolpert. “Bayesian optimization with a finite budget: An approximate dynamic programming approach”. In: *Advances in Neural Information Processing Systems* 29 (2016).

[16] Marco Laumanns and Jiri Ocenasek. “Bayesian optimization algorithms for multi-objective optimization”. In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2002, pp. 298–307.

[17] Jonas Močkus. “On Bayesian methods for seeking the extremum”. In: *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*. Springer. 1975, pp. 400–404.

[18] Carl Edward Rasmussen. “Gaussian processes in machine learning”. In: *Summer school on machine learning*. Springer, 2003, pp. 63–71.

[19] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. “A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions”. In: *Journal of Mathematical Psychology* 85 (2018), pp. 1–16.

[20] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical Bayesian optimization of machine learning algorithms”. In: *Advances in Neural Information Processing Systems* 25 (2012), pp. 2951–2959.

[21] Jasper Snoek et al. “Scalable Bayesian Optimization Using Deep Neural Networks”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 2171–2180.

[22] Rainer Storn and Kenneth Price. “Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces”. In: *Journal of global optimization* 11 (1997), pp. 341–359.

[23] Yanan Sui et al. “Stagewise Safe Bayesian Optimization with Gaussian Processes”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 4781–4789.

[24] Scott Sussex, Anastasiia Makarova, and Andreas Krause. *Model-based Causal Bayesian Optimization*. 2023. arXiv: 2211.10257 [cs.LG].

[25] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge, 1998.

[26] Ky Khac Vu et al. “Surrogate-based methods for black-box optimization”. In: *International Transactions in Operational Research* 24.3 (2017), pp. 393–424.

[27] James T Wilson et al. “The reparameterization trick for acquisition functions”. In: *arXiv preprint arXiv:1712.00424* (2017).

[28] AG Zhilinskas. “Single-step Bayesian search method for an extremum of functions of a single variable”. In: *Cybernetics* 11.1 (1975), pp. 160–166.

Study design for human acuity in symbol recognition

Oliver Veitl

Bayrische Motoren Werke AG
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
oliver.ov.veitl@bmw.de

Abstract

To date, there are no adequate quality parameters for see-through displays based on augmented and virtual reality especially for future automotive applications available. Recent approaches define detected stimuli in sinusoidal grids by their size, spatial and temporal frequency contrast sensitivity, luminance and eccentricity. The approach in this work is to know the background luminance and the display luminance distribution in order to define the contrast local rather than global for displays. This approach is based on the assumption that the ambient luminance distribution has a major influence on human visual acuity and its parameters. Therefore, a quantitative study concept is proposed based on a case study and the derivation of the relevant parameters.

By means of the described investigations it is possible to define the operating range of the optical human-machine interaction and the relevant optical parameters in head-up displays.

1 Introduction

One of the most volatile research topics of the last decade has been head-mounted and head-up displays. These displays provide additional information to the user via their visual stimulus. Acceptance of a new system depends on the quality delivered. The term quality is defined by the International Standards Organisation, short ISO, as the ability of a delivered product to satisfy all desirable and undesirable stakeholders. The relevant effects in head-up displays are optical defects.

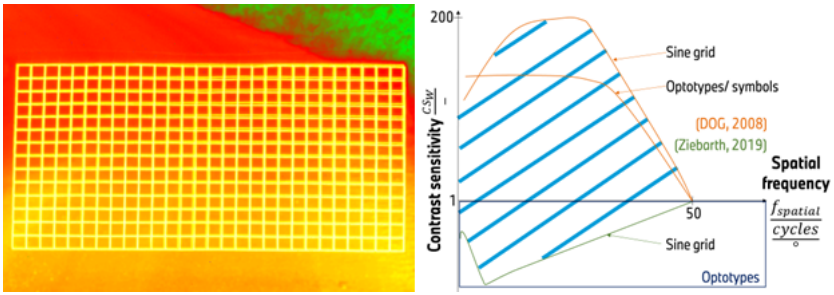


Figure 1.1: left: Luminance distribution perceived by the user of a head-up display.; right: Known fields of recognizable effects in various surroundings, see [3], [19]

Two of such effects are shown in the left part of Figure 1.1, there one can see the defined optical defects distortions and ghost images. Later can be recognized in several doubled lines in the upper right corner. It is defined as the second image detected in the direct vicinity of the first image. The order of the two images is defined by the perceived luminance difference between them, the first one is brighter than the second one. The mentioned optical distortions are defined as a deviation from an expectated line. Which can be observed in the middle of the top line. The distortion can be identified at the point of the local ghost image within the run of the line from left to right.

These effects can be seen at several points in the left part of Figure 1.1. We can now ask, what kind of defects in optotypes or symbols can the human visual system detect when interacting with head-up displays? To answer this question,

it is necessary to test human visual acuity in a generalised psychophysical study. In this paper we present a possible study structure for human visual interaction with high-luminance displays.

The meaning can be derived from the right part of Figure 1.1, which shows the relation between the perceived effect and the user's environment. The environment as the contrast sensitivity ordinate is applied over the recognized spatial frequency as the abscissa. The studies carried out show a clear difference between the recognition of sinusoidal patterns and symbols within the same environment. Different environments are investigated for the recognition of sinusoidal patterns. In addition, the influence of different viewing distances in different environments is an unknown parameter. This explains the white spot of the knowledge for this research. In order to investigate the influence of adaptation and accommodation effects on symbol recognition in human vision in the context of user interaction with high luminance displays, a generalised study structure is required with respect to the relevant parameters.

In order to bring the reader up to speed, the relevant literature research results are presented in Section 2. This is followed by the preliminary remarks in Section 3 to understand our proposal in Section 4. The paper concludes with a discussion of our findings and future work.

2 Related Work

As seen in the right part of Figure 1.1, during the decade of 2010, several studies have been conducted on human pattern recognition, see [3], [19]. Der Ophthalmologe published two studies showing the difference between optotypes and sinusoidal gratings. To obtain this data, the researcher used a standardised measurement system that allows a correlation to be made between contrast sensitivity, human visual acuity and recognisable sine grids or optotypes in different environments, see [3]. The background luminance is constant during the variation of the stimulus luminance. The values and structure of the study are based on standardised procedures [11]. In parallel, a research team from Cambridge is investigating sine grating detection and possible quantification parameters with their virtual model of the human visual system, see [13]. The model consists of the comparison of two different images. The image is manipulated using parameters and formulas developed from a number of different empirical studies. A recognised image has known defects and the effects of the human visual system are added, such as neural noise and magnification functions. The latest results of this research group show a lack of knowledge in several areas. Firstly, the model is generated by different studies with different objectives, [13]; [14], secondly there are unknown parameters such as sufficient accommodation of the human eye, see [8]. The studies shown all focus on optical distortions.

These results show a scientific need for a new data set. This should include the adaptation and accommodation variance for high and low luminance displays. In addition, to fill the gap in Figure 2.1, studies should focus on optotypes and symbols. To close the gap between the studies we have shown so far and the topic of head-up displays, we will look at the other side of the recent findings.

The first study to take into account the aberrations of the mirror surface was carried out by Neumann. He discusses different optical distortions by grouping them into local and global distortions. The difference between these groups is the area of effect. This means that a global effect is seen on the entire image, as in Figure 1.1. A local effect covers a much smaller area in relation to the global effect, the difference will be clarified in Section 3, see [16]. With regard to the findings of the research team from the University of Cambridge, another

hypothesis can be made. The local distortions are optical waves with a high spatial frequency, while the global distortions have a low spatial frequency. Considering the influence of the mirror surface on the reflected image, one of the summary statements in Neumann's paper, the link to Ziebarth's research becomes clearer.

The main results of this work show the correlation between the detection of an optical defect in sinusoidal gratings and its cause. The mathematical definition of the detectable optical distortions by the contrast sensitivity function is a suitable method, as described in the previous sections. To find the cause of the distortion, the specular flux is calculated for two different frames. The conclusions show the mathematical correlation between the optical spatial frequency and the geometrical spatial frequency, see [19]. In a parallel research work, the researcher investigated the quantification of optical distortions based on convex transparent specular surfaces. In addition to the previously presented work, recognisable parameters were defined. The study examines optotypes such as the grid shown in right part of Figure 2.1. The parameters symmetry, local distortion and shape can be identified by all subjects. The deflectometric measurements of the evaluated specimens are analysed in terms of their geometric properties. Based on this knowledge, virtual specimens are calculated and additionally evaluated by the same subjects. No correlation between geometry and optical distortion was found. Two assumptions could explain the results. Firstly, the ghosted fringe patterns, sinusodial grids reflected on transparent specular surfaces, do indeed have a relevant influence on the geometric distortions. Secondly, the design of the study needs to be much more specific to the hypotheses tested, see [2]. Another study concept is to test some quantification criteria towards the human visual system. The hypothesis is not confirmed during the analysis, see [18]. All three studies on reflected image quality have one parameter in common, an assumed constant lighting scenario and therefore a constant contrast sensitivity of the subjects. Aprojanz and Wagner do not take contrast sensitivity into account. While the latter focuses on the optically evaluated optotypes and defines a contrast ratio in a non-standardised form of $90 \cdot L_{Max}/L_{Min}$, [18] the former can describe the shape deformations based on the optical distortion of the optotypes, see [2]. Comparing Ziebarth and Wagner, the dynamic evaluation of the subject makes more defects visible, see [19]; [18].

As Mantiuk et al. show the relevant influence of contrast sensitivity and the unknown influence of accommodation, [15], [8] we propose a vision system and a study structure for the evaluation of sine grids center Figure 2.1, optotype grids left Figure 2.1 and symbols right Figure 2.1, taking into account the relevant parameters.

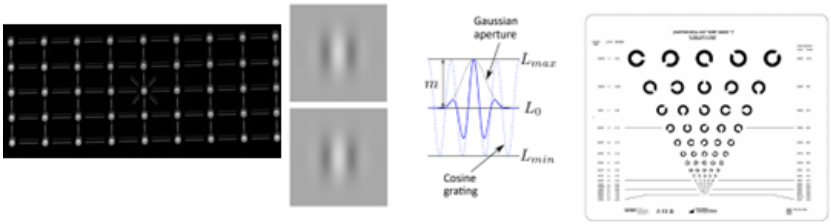


Figure 2.1: left: Tested optotype grids for head-up display quality, see [16].; center: Tested sine grids for perception values of contrast-sensitivity, see [14].; right: Tested optotype grids for precepted values of contrast sensitivity, see [17].

3 Psychophysical Relations

To understand the difference between the previous work and our new approach, in the following section we derive the model of human visual adaptation and accommodation for a high-luminance display. As can be seen in Figure 1.1, the user of such displays sees a specific image superimposed on the environment. This means that we claim that the human visual system is adapted to a certain environment, as seen in right Figure 3.1. After the superimposition, the eye accommodates to a given image, as seen in left Figure 3.1. Both of the following images are logarithmically scaled to give an idea of the actual predicted lighting scenario. We show a sunny and clear day scenario, taken with a calibrated luminance meter and a resolution of $2448 \times 2048 \cdot [Pixel \times Pixel]$.

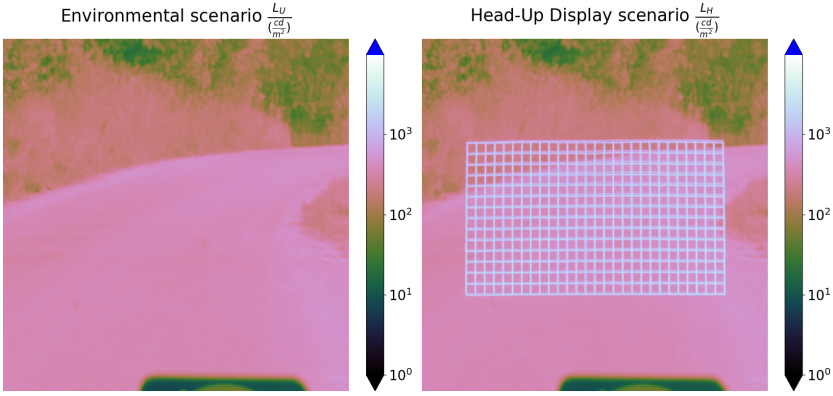


Figure 3.1: left: Adaption relevant luminance distribution.; right: Superimposed luminance distribution with accommodation relevant image.

As can be seen in the step between left Figure 3.1 and right Figure 3.1, the adaptation is independent of the superposition of the environment. This claim is based on subjective perception. If the additional luminance is influencing adaptation, the environment would become darker relative to the display luminance. Several subjective experiments have shown the same effect, i.e. the superposition of different luminance values does not affect the perceived environment.

The results in Section 2 assess this effect on the basis of the Michelson contrast. It is defined by the fraction of a luminance difference over the luminance superposition, as seen in Equation (3.1), see [14], [4].

$$C_M = \frac{\Delta L}{L} = \frac{(L_{max} - L_{min})}{(L_{max} + L_{min})} \tag{3.1}$$

The interpretation of the declarations in Equation (3.1), see [4], is shown in Figure 3.1. As we know from Section 2, since the results of Der Ophthalmologe, the contrast and contrast sensitivity in optotypes should be described by the Weber contrast function Equation (3.2), see [4]:

$$C_W = \frac{\Delta L}{L_{Max}} = \frac{(L_{max} - L_{min})}{L_{max}} \tag{3.2}$$

In this paper the put values are interpreted differently. The maximum value L_{Max} is the luminance at one point within the superimposed image, which we define as follows $[L_{HUD}] = \frac{cd}{m^2}$. The minimum value L_{min} is the measured value of the environmental luminance at the same point, defined as $[L_{Env}] = \frac{cd}{m^2}$, see [5]. Since the publication from Der Ophthalmologe the correlation between the visual acuity and the contrast sensitivity Equation (3.3) is known, see [5]:

$$CS_W = \log_{10}(\frac{1}{C_i}) = \log_{10}(\frac{1}{C_W}) \tag{3.3}$$

Contrast sensitivity in equation (3.3) is defined as $CS_W = [\frac{1}{\%}]$. This function is an essential value to describe the irradiance differences between two stimuli on the human retina. The interpretation of whether a human can recognise two stimuli or one stimulus is defined by the Rayleigh or Sparrow criteria. The application in Figure 3.1 illustrates this. The criteria are defined on the basis of the prescribed wavelength of the light and the aperture, in our case the human iris, of the maximum visual acuity, see [9]. Summarized, this means, are two lines differentiable or not. Figure 3.2 shows the computation of the human visual potential for different wavelengths based on different apertures.

The Figure 3.2 shows the computation angular visual acuity for photopic or day vision and scotopic or night vision. To describe these two scenarios, the pupil

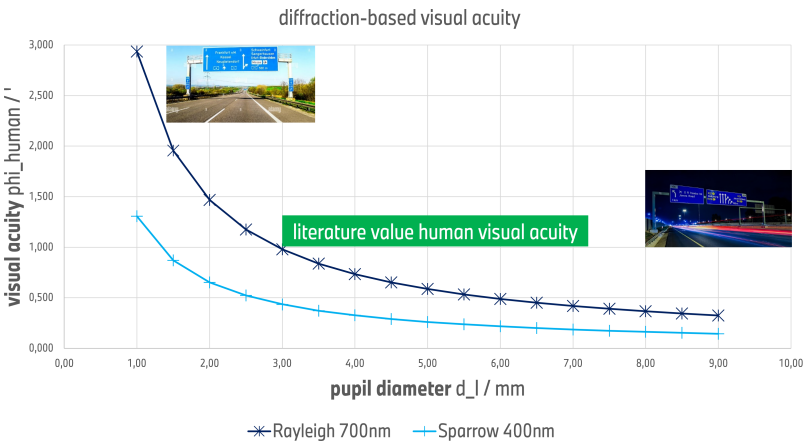


Figure 3.2: Computation of the human visual acuity in dependency of the pupil diameter for the Rayleigh criteria and the Sparrow criteria.

diameter based on the literature is used. The computation shows a possible set of solutions from $0,86 \cdot mrad > \varphi_{Human}(1,0 \cdot mm) > 0,38 \cdot mrad$ down to $0,10 \cdot mrad > \varphi_{Human}(9,0 \cdot mm) > 0,04 \cdot mrad$, the conversion from arc minute to radians is calculated by the constant $K_{mrad} = \frac{\pi}{(60 \cdot 180 \cdot 1.000)} \cdot \frac{mrad}{'}$. In comparison, the literature states the maximum visual acuity is defined as $\varphi_{Max}(Human) = 0,29 \cdot mrad$, see [10]. This angle is defined as a function of the nodal point position and the statistical distribution of the retinal cones. Therefore, a study design needs to consider the set of solutions and the environmental parameters are elaborated.

3.1 Human vision

Since the function of the human visual system is not trivial and the psychological parameters in perception are not negligible, one should first clarify the observed parameters. The parameters and the method of observation have an additional influence on the classification of the subject study. The upper and lower Figures 3.3 show two functions. The upper in Figure 3.3 is an abstraction of an ideal adaptation scenario. It is ideal because the optical imaging of the object onto the curved image plane is free of distortion and every average ray passes through the nodal point K_i'.

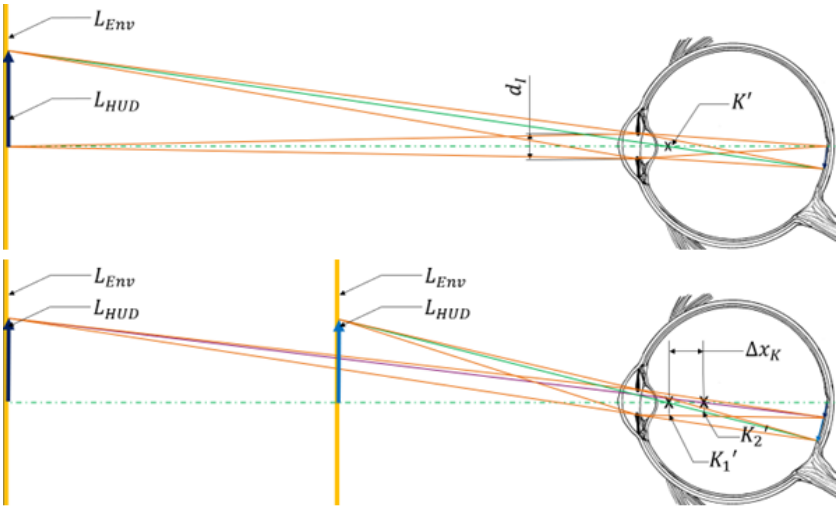


Figure 3.3: upper: Human vision adaption model for high luminance displays.; lower: Human vision accommodation model for high luminance displays.

Figure 3.3 shows the assumed adaptation model used for the study design. The human eye on the right side is based on the Gullstrand model see [4]. Its fixation point is a part of the object plane described by the luminance L_{HUD} and its orientation by the dark vector. The focused plane is surrounded by the yellow environment described by L_{Env}. Accommodation is a change in the iris diameter to limit the incoming light flux and is defined as $[d_i] = mm$. The accommodation model is abstracted in lower Figure 3.3. The Gullstrand eye model focuses on two different levels. The first is defined by the black vector starting from the optical axis and the second by the blue vector. The constant adaptation is achieved by an equal environment L_{Env}. The orientation defined by the vector and the size of the observed dark and blue image are radiated by an equal luminance L_{HUD}. An image of the same size is perceived as larger if the distance between the nodal point and the focal plane is smaller. In addition, the human eye can adjust the pupil to obtain a sharp image. A possible explanation is a longitudinal shift of the nodal point. The shift is then realised by the relaxation or contraction of the pupil. This effect results in different nodal points, K_1' for the near image and K_2' for the far image. In summary, the proposed study is based on the change in visual acuity, defined as $[\Delta\varphi_{Human}] = mrad$, by a variance in the adaptation-based pupil diameter d_I and the accommodation-based longitudinal nodal shift $[\Delta x_{longitudinal}] = mm$. A user may perceive different types of optical defects in the scenario shown in right Figure 3.1, distortions see left Figure 3.4, ghost images see center Figure 3.4 and disparities see right Figure 3.4.

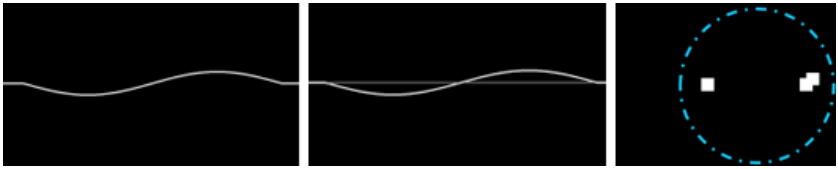


Figure 3.4: left: Distortion based on a cosine; center: Ghost image based on a cosine distortion; right: Disparitiy point shift

The distortions can be described by a wave. The mathematical definition of the functions shown is not relevant as it is only intended to clarify possible parameters. It becomes clear that amplitude, spatial frequency, contrast and

the experience of the observer influence the detection of distortions. This will be discussed in the following section. The ghost image is also influenced by the Rayleigh or Sparrow criteria already mentioned. The most complex are the disparities. Disparities are optical defects that affect the human stereoscopic vision system. Figure 3.4 (right) contains a blue circle to define the area under investigation. The area describes the spatial region in which the empirical horopter can be abstracted as a plane. The horopter is a virtual ellipse in the human environment, it describes points of equal depth estimation based on the triangulation of both eyes and the human image processing calculated by our brain. The two points are equally distant from the central plane of both eyes. On the right, there are two white shifted dots. This shift means that the points do not represent the matching cones of the right and left retinas. A new set of parameters is therefore relevant for human recognition, consisting of anthropometric data, in particular the interocular distance, and the Rayleigh or Sparrow criteria. To summarise the basics of human vision, the following parameters are considered.

$$\varphi_{Distortion} = f(d_I, x_K, \omega_{Distortion}, A_{Distortion}, CS_{W_i}) \tag{3.4}$$

$$\varphi_{Ghost} = f(d_I, x_K, \omega_{Ghost}, A_{Ghost}, CS_{W_i}) \tag{3.5}$$

$$\varphi_{Disparitiy} = f(d_I, x_K, \Delta x_{Disparity}, \Delta y_{Disparity}, CS_{W_i}) \tag{3.6}$$

Equation (3.4) describes the visual acuity $\varphi_{Distorion}$ for distortion detection in optotypes, it is a function defined by iris diameter d_I, nodal point position d_K, the spatial frequency $\omega_{Distorion}$, the defect size $A_{Distorion}$ and the contrast sensitivity CS_{W_i} between the environment and the image plane. While the detection of a ghost image depends on the same parameters, the contrast sensitivity reference additionally takes into account the irradiance of the second line in left Figure 3.4. The mathematical parameters of disparities are described by a longitudinal shift $\Delta x_{Disparity}$ and a lateral shift $\Delta y_{Disparity}$. The contrast sensitivity in the case of disparities describes the ratio between the background and the two shifted dots on the right side of right Figure 3.4. The set of new parameters needs to be considered in a single study, a properly structured study and the consideration of psychological effects, see [19]; [14].

3.2 Psychological study design

The research process of empirical investigation of psycho-physical relations follows a strict procedure. For quantitative results with high conceptual, internal, external, and statistical reliability the sequential process includes 9 steps, see [6].

1. Research topic
2. State of the art
3. Study design
4. Operationalisation
5. Sampling of subjects
6. Execution
7. Study processing
8. Analysis
9. Presentation

Steps one and two are covered in the first two sections of this paper. This chapter focuses on the third step in the upper list. In order to classify the required study, the hypotheses to be investigated are formulated at this point.

- Is there a significant difference in human visual acuity for contrast sensitivities below 1 for symbols and optotypes compared to standardised methods?
  - H_0: There is no statistically relevant difference between the concept and standardised methods.
  - H_1: The human visual acuity is statistically lower than standardized methods.

- Is there a dependence on human visual acuity for optotype defect recognition with respect to the independent variables investigated (defect size, accommodation and adaptation)?
  - H_0: No dependency function between the independent and dependent variable is found.
  - H_1: The variables dependent and independent show a relation between each other.

The hypotheses show two expected quantitative outcomes. The outcome studied or measured in psychological experiments is called the dependent variable. The varied parameters are called independent variables. Confounding variables may also influence the desired outcome. In the study design, all known variables are classified, the execution plan is structured, and the analysis method is defined taking into account the required validity, see (Döring, 2022). The validity of the study design depends on its internal and external validity and its methodological rigour. The classification of studies into the two types is helpful.

1. Research-theoretical approach
2. Knowledge objective
3. Object study
4. Data basis
5. Knowledge interests
6. Subject line-up and treatment
7. Investigation surroundings
8. Investigation repetitions (time)
9. Investigation repetitions (subject)

The desired measurement outcome influences the chosen approach by its predefined scale. If the output is unknown to the scientific community, the knowledge objective is basic research, otherwise it is called an applied research study. A study is defined by the objects studied, whether it is theoretical research in articles, methodological research for further development, or gaining new knowledge through empirical studies. This makes it possible to define the class of data needed. Is the data set generated new, so-called primary data, are there already secondary data sets that can be analysed with new methods, or is it possible to combine several existing data sets statistically for a meta-analysis? This may be done for various reasons. One reason may be an exploratory study to derive research questions or hypotheses, an explanatory study to prove pre-defined hypotheses as possible answers to open questions, or descriptive studies to test hypotheses for a representative entirety. The population to be studied is also part of the sample and its treatment. A possible subject line-up affects the quality of the data provided. To investigate a causal relation, it is recommended to choose an experimental study with randomised groups. There is no control on the group composition, but the effect can be measured by the installation of a control group. The control group is an instrument to elaborate cause and effect by variation of effect between the groups. A requirement for this installation is a parity of both groups. This method is also used in quasi-experimental studies with the difference of manipulating the group composition. Another concept that is possible are non-experimental studies. The groups are not randomized and the potential for causal investigations is low. Another important issue is the setting. Typically, designs are described as a field or laboratory study. The latter is a simulation of a real situation, while the former is measured in a real situation. The difference between these two approaches is the controllability of the confounding variables. This is where internal and external validity become more important. Both define the quality of the study results. An internally valid result provides a clear causal relation between the dependent variable and the independent variable studied. Therefore, methods such as randomisation for subject confounding and constant environmental confounding with proper control and documentation are used. Externally, a dependent variable is valid if the design and the whole are generalisable. This can be provided by a representable group of subjects. Another possibility is to specify the subject by their ability to

understand the changes during the study. These levers allow a lower variance of the dependent variable. Replication also reduces variance and increases validity. Time-based repetitions describe the use of subjects. They take part in either a within-subject study or a between-subject design. The latter omits repetitions to save time, while the former includes a preliminary measurement to increase internal validity. Within-subject designs have additional problems, depending on the particular case. One problem may be the acquisition of expertise. This is especially the case when the subject has to perform a certain task under different conditions. The time available for the task may decrease with each attempt. The final point is the repetition of tests by increasing the number of subjects. The absence of replication means that a case study has been carried out. In scientific work it is more common to carry out group studies, see [6]. In order to find a causal relation between the dependent variable and the independent one over a representative entirety, two additional points have to be clarified, the statistical analysis method and the classification of the single measurements. One single run in a conducted psychophysical study is classified by three parameters, see [12]. The first level of description is the class, followed by the dichotomous level. The first is defined by the absence of parallel stimuli. This means that a reference element can be used by the subject. Other dichotomies used to describe an experiment are decision types, evaluation types and task types, see [12]. The decision for an answer to a question can be a forced choice or a rating on a given scale. The questions relate to the subject's evaluation, such as identifying a difference or absolute steps, or detecting errors relative to a reference versus finding the worse error in discrimination tasks, and finally the objective or subjective measurement of representations. These assessment tasks are subdivided into the measurement groups of magnitudes or dimensions, called appearance, and performance measurements, related to the subject itself, see [12]. The analysis of the individual runs and the overall study can be carried out using different statistical methods in relation to the tasks. Döring emphasises the prior knowledge required for a "correct" measurement, see [6].

3.3 Statistical distributions and tests

In the last chapter the relationship between hypotheses and questions was mentioned. To test hypotheses, statistical methods are needed. The choice of the right methods depends on the data set and the idea of the possible distribution. In this section we will look at two different distributions and the mathematics of testing them. The first is a two-sided distribution, also known as a symmetric distribution, and the second is a one-sided or two-sided distribution. The latter can be described by the Chi-square method and the former by a normalisation method. Finally, a coefficient is introduced to prove a possible correlation between different variables. An example of a two-sided distribution is the normal density function. The function is symmetrical and describes the density of samples at a given value. The variance of the samples tested around a given mean or median and their density of occurrence is described by a function, see Figure 3.5. Three different distributions are shown in the figure, see [7]. The graph shows the density of the tested samples over their initial value. The dotted function has a lower variance of the output values on the abscissa and therefore a higher density of values around the mean. The dashed line, on the other hand, shows a high variance and therefore a low density of values in the range of values on the ordinate.

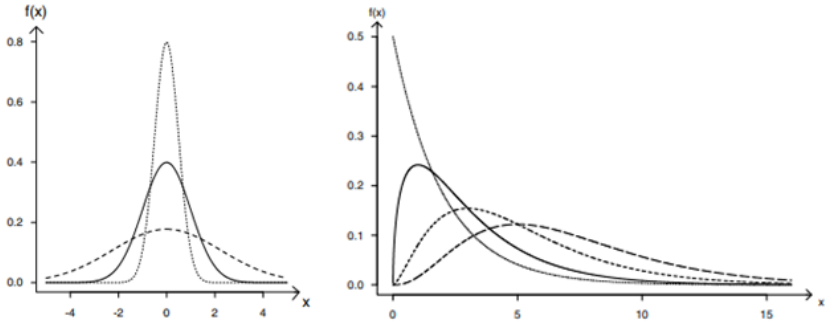


Figure 3.5: left: Normal density function, see [7].; right: Chi-squares density function, see [7].

Right Figure 3.5 shows a chi-squared density distribution for different samples tested. This function is a permutation of the normal distribution in left Figure 3.5, see [7]. The median shift is recognizable between the solid, the dashed and the dotted lines from left to right. The density of values for the specific functions shrinks in the same direction, causing the spread function to increase. A left boundary is the reciprocity of all functions in right Figure 3.5. Testing the sample distribution against a chi-square function is part of the cross-correlation tables. The reliability of the median depends on the number of samples. The ideal case is testing $N_{Samples} > 50$, a mathematical error correction is only possible for $N_{Samples} \geq 20$. Starting from the correct number of samples, it is possible to analyse a significant difference between the observed and the expected value, as shown in Equation (3.7):

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h_{ij}})^2}{\tilde{h_{ij}}} \tag{3.7}$$

The test coefficient χ^2 for an $i \times j$ dimensional measurement matrix is computed by the squared distance function between the observed statistical frequencies h_{ij}, and the expected one $\tilde{h_{ij}}$ as a fraction over the expected statistical frequency $\tilde{h_{ij}}$ The function is used in various test procedures, independency test, distribution test, and homogeneity test.

$$df = (Ni - 1) \cdot (N_j) \tag{3.8}$$

To evaluate the significance, the degree of freedom is needed, computed by Equation (3.8). It takes into account the number of discretisation steps in the test matrix for $i \times j$. The computed value compared to the chi-square distribution table, see [7], provides the upper bound depending on the number of samples. As already mentioned, reliability is a strong evaluation method, therefore Equation (3.9), see [7] is a possible tool:

$$CramersV = \sqrt[2]{\frac{\chi^2}{\chi^2 + N_{Samples}}} \tag{3.9}$$

Cramérs rule or the contingency coefficient makes it possible to calculate the meaningfulness of the experiment, it should be bigger than $CramersV > 0,3$. Another possibility to analyse data is correlation analysis. Therefore, the samples are tested in multidimensional density rooms, as seen in Figure 3.6.

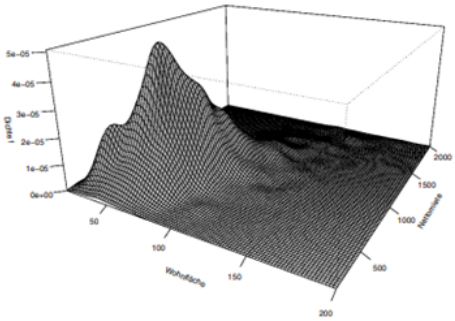


Figure 3.6: Two-dimensional density function, see [7].

The Figure 3.6 shows the density of samples over two different parameters. The computation of such a landscape is done by Bravais-Pearson (3.10) or Spearman (3.11) coefficients, see [7]:

$$r_{Pearson} = \frac{\sum_{n=1}^{k}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt[2]{\sum_{n=1}^{k}(x_i - \bar{x})^2 \times \sum_{n=1}^{k}(y_i - \bar{y})^2}} \tag{3.10}$$

$$r_{Spearman} = 1 - \frac{6 \cdot \sum_{n=1}^{k}(x_{rank} - y_{rank})^2}{n_{Samples} \cdot (n_{Samples}^2 - 1)} \tag{3.11}$$

Pearson's distribution can handle of non-linear functions, whereas Spearman's cannot. Spearman transforms the measured values of parameter one x and parameter two y of the scatterplot into a ranked list.

Whereas Pearson calculates the fraction of the summed product of the distance functions over the square root of the square sum of the individual distance functions of x and y with respect to their mean values $\bar{x}$ and $\bar{y}$. The statistical interpretation is the same for both distributions, see [7]:

$$t_i = \frac{r_i \cdot \sqrt[2]{n_Samples - 2}}{\sqrt[2]{1 - r_i^2}} \tag{3.12}$$

A significance t_i based on the distribution of Pearson p_i, equally Spearmann s_i, is given for squared values bigger $0,1$. In the ideal case $t_i^2 > 0,5$. Based on knowledge of statistical analysis methods, the psychophysical data acquisition and the optical conditions in the following section a study concept is derived. It considers correlation and independent variables.

4 Results

This section presents the structure of the study and the individual experiments. Relevant variables are derived from a preliminary field case study, the analysis defines boundaries and conceptual parameters. The derivation of the variable structure is based on the state of the art and the knowledge gained from the case study. Structured dependencies and variables lead to the proposed study design and experiments, taking into account the state of the art. The necessary hardware setup is then derived to investigate the variables in an environment similar to the field case.

4.1 Preliminary field case studies

This study elaborates on the variables and their limitations as mentioned in Section 3. In addition, the findings from Section 2 are taken into account in the design. During a sunny and clear day, typical user interaction scenarios are captured. Weather conditions during the night are also clear. To simulate the most accurate scenario, a tarmac road in a north-south direction is chosen. This is done to maximise the variance of the background luminance and to minimise unacceptable stray light affecting the measurement. The test is carried out on a production vehicle. Unknown parameters, see Section 3, are the adaptation-luminance relationship and accommodation. As the latter cannot be measured without violating the system under investigation, only the former is shown in Figure 4.1.

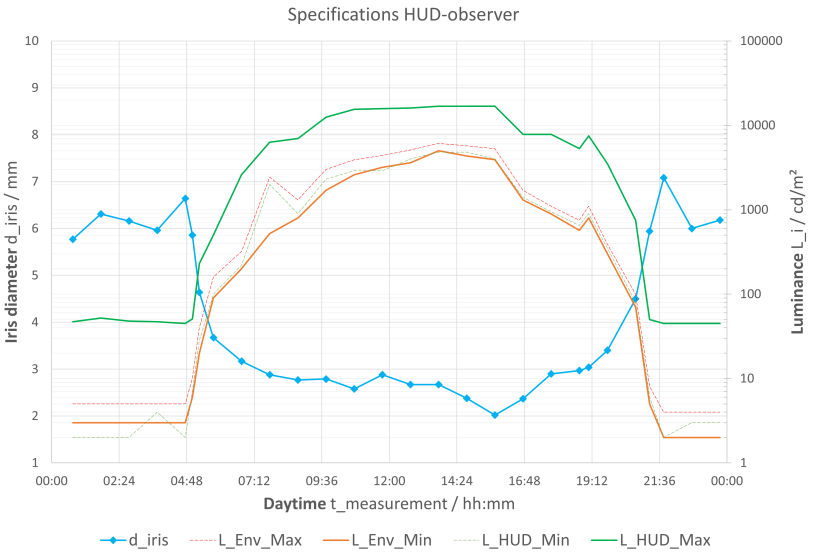


Figure 4.1: Caputered relation between the iris diameter and the lightning scenario in high-luminance displays.

The upper graph shows the specifications for human vision as defined in Section 3. Over a 24 hour period, the measured data record the diameter of a human iris on the primary ordinate and the perceived absolute luminance extremes on the secondary ordinate. The extremes are taken from the distributions shown in left Figure 3.1 and right Figure 3.1. The constant distribution at night is caused by a constant background distribution from car headlights. This results in a constant aperture diameter, taking into account the measurement uncertainty. Another effect is the adaptation of the minimum ambient luminance caused by the head-up display system. The system superimposes a grid on this distribution, as shown in right Figure 3.1. A peak in the aperture diameter is measured at sunrise and sunset. No valid explanation can be derived from the data shown. A possible hypothesis is the high level of blue light during the 'blue hour' or sunrise and the high level of red light during the 'golden hour' or dawn.

No further investigation is carried out at this point, only the maximum iris diameter during driving scenarios of $d_I = 7 \cdot mm$ is noted as a control variable. From this point on, the diameter decreases towards the base ten in the same way as the logarithmic luminance distribution. Between 8 o'clock and 18 o'clock, the diameter remains constant at a median value of $d_I = 2,7 \cdot mm$, taking into account the measurement uncertainty. The absolute minimum at 15 o'clock is disregarded due to its single occurrence and the almost constant luminance distribution over several hours, a further investigation should clarify the event. Upper boundary maxima are considered for their double appearance. Diameter-boundaries are reached at the luminance differences of $L_{Env} = 7 \cdot \frac{cd}{m^2}$ to $L_{Env} = 45 \cdot \frac{cd}{m^2}$ and $L_{HUD} = 900 \cdot \frac{cd}{m^2}$ to $L_{HUD} = 7.000 \cdot \frac{cd}{m^2}$. These situations and boarders are noted as possible boarders and need therefore another analysation towards the optical parameters mentioned in Section 3, as seen in Figure 4.2 and Figure 4.3.

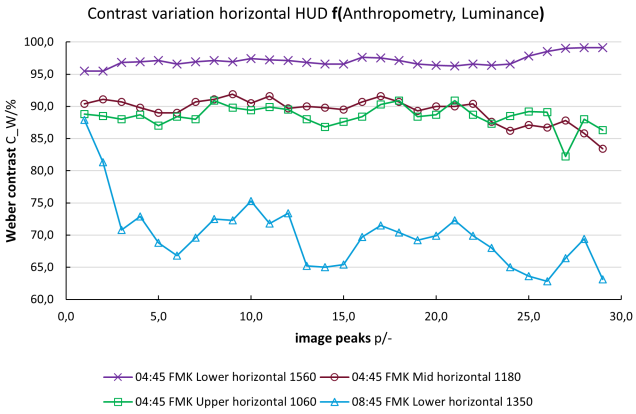


Figure 4.2: Analysation of the extremes of Figure 4.1 towards the human perceived contrast.

The discussion of Figures 4.2 and 4.3 must be done in parallel to see the relevant parameters. The ordinates show a section through the grating, see right Figure 3.1, for the calculated contrast and contrast sensitivity. Each measurement point represents a peak of the grating.

Four representative situations were chosen to define the limits. The first three data sets representing the different driver sizes are dyed in purple, green and red. The fourth one is chosen to represent a group of data sets, which are varying arround a lower contrast limit of $C_W = 70 \cdot \%$. The anthropometric data, taking size into account, shows that the upper limit in terms of contrast is represented by a large driver. In terms of contrast sensitivity, which is related to human visual acuity, this is the lower limit. The upper limit is measured for the same driver at high luminance. Large internal changes in contrast distributions can be correlated with changes in background luminance.

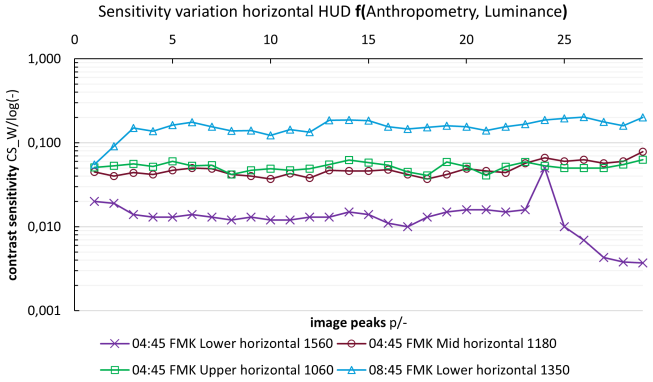


Figure 4.3: Analysation of the extremes of Figure 4.1 towards the human perceived contrast sensitivity.

The range of measurement data in Figure 4.3 proves the proposal of Section 1 towards the investigation area. Boundaries defined based on the preliminary study are $0,0037 < CS_W < 0,2$. Summarizing the knowledge of the case study and the state of the art the study structure can be set.

4.2 Psycho-physical sample study concept

In the first step, see Section 3, the relevant variables are summarized and classified, see Table 4.1.

Name	Var.	Treatment	scale
Iris diameter (eye)	CV	Environment	Preliminary
Visual acuity I (eye)	IV	DIN EN ISO 8596	Preliminary
Visual acuity II (eye)	IV	Environment	Processing
Pattern orientation	DV	Horizontal/vertical	1-2
Frequency (pattern)	DV	Randomized	1-5
Amplitude (pattern)	DV	Randomized	1-5
Offset (Pattern)	DV	Randomized	1-5
Shift (Pattern)	DV	Randomized	1-5
Distance (pattern)	DV	Grouped	1-9 m
Contrast sensitivity	CV	Grouped	Preliminary

Table 4.1: Elaborated parameters from the preliminary study to be considered in the study construct.

Table 4.1 groups the relevant known parameters according to their role in the study. The independent variables, short IV, iris diameter and visual acuity, are measured during the study. These are actively manipulated by the dependent variable, short DV, and the confounding variable, short CV. As can be seen in the table, the diameter is not only a CV, but also contrast sensitivity. These variables are documented during the study to exclude singularities. Another point is the detection of fatigue symptoms by recording the diameter. Experience shows that the human visual system requires a great deal of effort to focus on high-luminance displays over time, this record makes it possible to correlate drifts within an experiment with symptoms. The third variable is measured first to falsify the first null hypothesis. The third IV is the reference visual acuity, which is measured using standardised methods, see [11]. The perceived luminance distribution is recorded to match the quantitative values in the post-processing. These values are stored for all known optical defects, see Figure 3.4. The manipulation of

horizontal or vertical lines and dots is selected in this step. As every human has a high level of experience with these optotypes, the visual system is additionally 'calibrated' to an intrinsic vertical and horizontal line, see [4]. In summary, a 3×3 study structure is required in the lighting scenarios already defined, see Figure 4.4. The study shown provides a 3×3 measurement matrix for each of the optical defects. The matrix is analysed using the correlation methods already mentioned. Each point in the matrices is obtained by a 3×5 experiment in laboratory conditions derived from the field studies described in the last chapter. As discussed in Section 2, the randomisation of stimuli and their repetition is essential for high internal validity.

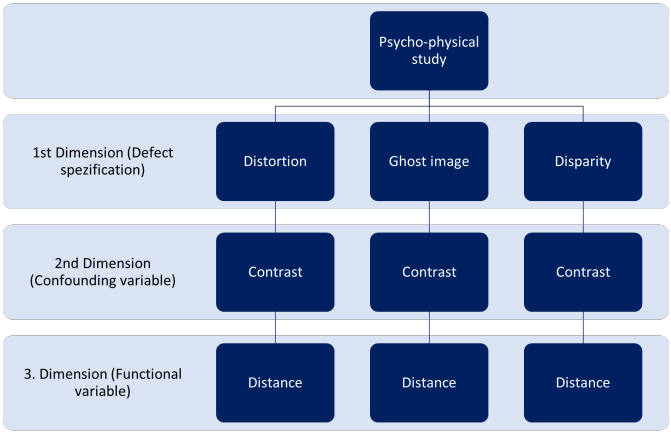


Figure 4.4: Summarized study structure for a 3×3 factor plan considering all adaption and accommodation effects.

In order to minimise the empirical and temporal effort, only the detection amplitudes for the different scenarios are divided into 5 steps. It is assumed that at least three grid points are needed to distinguish between a linear and a non-linear relation. As mentioned in Section 2, the frequency and amplitude of the defects in the patterns are decisive for the distortions. In the ghost image, the measuring points are defined by the offset between two lines and their contrast sensitivity to each other. The last type of defect is disparity, in which case two points are shifted relative to each other and their contrast sensitivity is defined, as shown

in Figure 3.1. In order to see a significant difference, the standardised acuity is measured and each value of the 3×5 contingency matrix obtained from the experiment in Figure 4.5 is analysed by the Chi-square method.

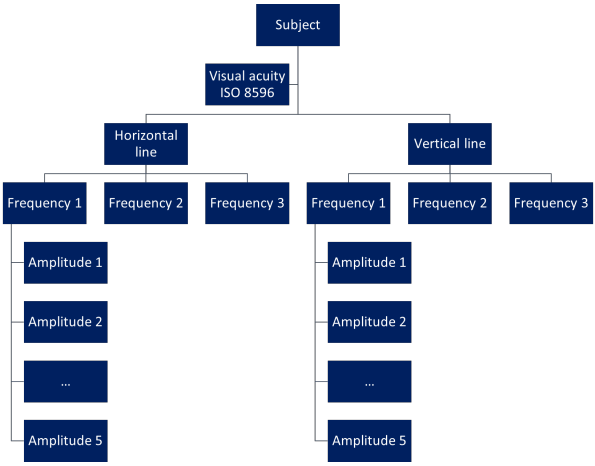


Figure 4.5: Sample experimental plan to derive the visual acuity in optotypes in comparison to standardized methods.

Before visual acuity is measured, the subject must adapt their visual system to the environment. The maximum adaptation time for photopic vision is measured as a period of ten minutes, see [1]. After the aforementioned adaptation period and acuity measurement in the final environment, full adaptation should be allowed for the forced-choice viewing of 90 images in 15 minutes.

5 Conclusion and further work

The paper presents an empirical concept based on the assumption of the importance of the local contrast sensitivity. The explanatory objectives are achieved by an optical quantitative data set derived from laboratory position experiments. The laboratory description of the parameters of the optical field study increases the external validity, whereas the internal validity is increased by the repetition of the ratings by the subject and the repetition of the experiment by different subjects. This methodology increases the statistical validity. This approach provides a complete set of primary visual acuity data. Bottle necks of the methodology can be the chosen statistical post-processing methods. Analysis of such data sets using chi-squared distributions has already been carried out, but from a critical point of view the distribution of the data is not yet known. Knowing this, we only can assume a one-sided bounded distribution, which can also be described by a log-normal distribution. The proposed optotypes also have an effect. In the case shown, lines are assumed to have a strong influence on perception. The discussion of the results is concluded by mentioning the lack of a method to control the subject's state of accommodation. A setup that fully meets the optical requirements listed in Table 4.1 is shown in Figure 5.1.

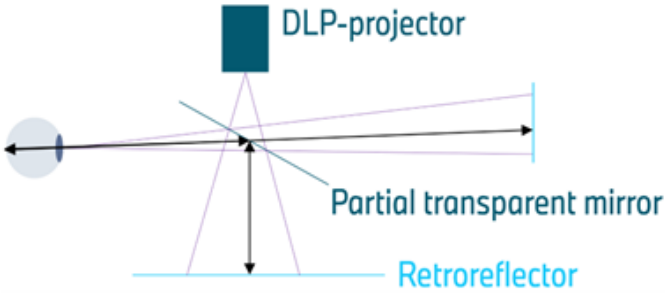


Figure 5.1: Sample experimental plan to derive the visual acuity in optotypes in comparison to standardized methods.

In Figure 5.1, the adaptation- and accommodation state can be fitted towards the research plan, see Figure 4.4, also the experimental runs, see Figure 4.5, are feasible. Further reseach will be conducted based on the already known variables, see Table 4.1. The state of the art shows the complexity for such expirements, see [8]; [14]; [19]; [3], of the human visual acuity and the obtained measurment outcome. To meet these challenges, the spatial frequencies and contrast sensitivities received by humans, as well as the theoretical perceptual amplitudes, offsets and shifts, have to be detailed. In addition, future work will propose a geometric setup that takes into account the requirements studied.

References

[1] K.R. Alexander. “Information Processing: Retinal Adaptation”. en. In: *Encyclopedia of the Eye*. Elsevier, 2010, pp. 379–386. ISBN: 978-0-12-374203-2. DOI: 10.1016/B978-0-12-374203-2.00201-3. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780123742032002013 (visited on 11/07/2023).

[2] Stephanie Aprojanz. *Untersuchung zur Objektivierung... durch deflektometrische Messtechnik*. de. Wiesbaden: Springer Fachmedien Wiesbaden, 2019. ISBN: 978-3-658-24369-2 978-3-658-24370-8. DOI: 10.1007/978-3-658-24370-8. URL: http://link.springer.com/10.1007/978-3-658-24370-8 (visited on 11/07/2023).

[3] M. Bach et al. “Photopisches Kontrastsehen: Örtliche Kontrastempfindlichkeit”. de. In: *Ophthalmologe* 105.1 (Jan. 2008), pp. 46–59. ISSN: 0941-293X, 1433-0423. DOI: 10.1007/s00347-007-1605-y. URL: http://link.springer.com/10.1007/s00347-007-1605-y (visited on 11/07/2023).

[4] *DIN 5340*. Standard. Genf, Schweiz, 2022. DOI: 10.31030/3371845. URL: https://www.beuth.de/de/-/-/356879010 (visited on 11/07/2023).

[5] e.V. DOG. “Prüfung des Kontrast- oder Dämmerungssehens”. de. In: *Ophthalmologe* 108.12 (2011), pp. 1195–1198. ISSN: 0941-293X, 1433-0423. DOI: 10.1007/s00347-011-2488-5. URL: http://link.springer.com/10.1007/s00347-011-2488-5 (visited on 11/07/2023).

[6] Nicola Döring. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. ger. 6., vollständig überarbeitete, aktualisierte und erweiterte Auflage. Lehrbuch. Berlin [Heidelberg]: Springer, 2022. ISBN: 978-3-662-64762-2 978-3-662-64761-5. DOI: 10.1007/978-3-662-64762-2.

[7] Ludwig Fahrmeir et al. *Statistik*. Springer-Lehrbuch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. ISBN: 978-3-662-50371-3 978-3-662-50372-0. DOI: 10.1007/978-3-662-50372-0. URL: http://link.springer.com/10.1007/978-3-662-50372-0 (visited on 11/07/2023).

[8] Dounia Hammou. “Comparison of metrics for predicting image and video quality at varying viewing distances”. English. In: *IEEE MMSP* (2023). URL: https://www.cl.cam.ac.uk/~rkm38/pdfs/hamond2023_view_dist_metrics.pdf.

[9] Eugene Hecht. *Optik*. ger. Trans. by Karen Lippert. 8. Auflage. De Gruyter Studium. Berlin Boston: De Gruyter Oldenbourg, 2023. ISBN: 978-3-11-102525-4.

[10] Ekbert Hering and Rolf Martin, eds. *Optik für Ingenieure und Naturwissenschaftler: Grundlagen und Anwendungen: mit zahlreichen Bildern, Tabellen, Beispielen*. ger. München: Fachbuchverlag Leipzig im Carl Hanser Verlag, 2017. ISBN: 978-3-446-44281-8.

[11] *ISO 8596*. Standard ISO 8596. Genf, Schweiz: Beuth Verlag, 2018. DOI: 10.31030/2744962. URL: https://www.beuth.de/de/-/-/278501784 (visited on 11/07/2023).

[12] Frederick A. A. Kingdom and Nicolaas Prins. *Psychophysics: a practical introduction*. Second edition. OCLC: ocn946067924. Amsterdam: Elsevier/Academic Press, 2016. ISBN: 978-0-12-407156-8.

[13] Rafał Mantiuk et al. “HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions”. en. In: *ACM SIGGRAPH 2011 papers*. Vancouver British Columbia Canada: ACM, July 2011, pp. 1–14. ISBN: 978-1-4503-0943-1. DOI: 10.1145/1964921.1964935. URL: https://dl.acm.org/doi/10.1145/1964921.1964935 (visited on 11/07/2023).

[14] Rafał K. Mantiuk, Maliha Ashraf, and Alexandre Chapiro. “stelaCSF: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area”. en. In: *ACM Trans. Graph.* 41.4 (July 2022), pp. 1–16. ISSN: 0730-0301, 1557-7368. DOI: 10.1145/3528223.3530115. URL: https://dl.acm.org/doi/10.1145/3528223.3530115 (visited on 11/07/2023).

[15] Rafal K. Mantiuk, Dounia Hammou, and Param Hanji. “HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content”. In: (2023). Publisher:

arXiv Version Number: 1. DOI: 10.48550/ARXIV.2304.13625. URL: https://arxiv.org/abs/2304.13625 (visited on 11/07/2023).

[16] Alexander Neumann. “Simulationsbasierte Messtechnik zur Prüfung von Head-up Displays”. Deutsch. PhD thesis. Technische Universität München, 2012. URL: https://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20120227-1079689-1-0.

[17] Visus. *Visustafel Landoltringe*. Nov. 2023. URL: https://www.visus.de/media/image/ab/ef/09/81685_600x600.png (visited on 11/07/2023).

[18] Daniel Wagner. “Wahrnehmung von Augmented Reality Head-up Displays”. de. In: (2023). Medium: PDF Publisher: Karlsruher Institut für Technologie (KIT). DOI: 10.5445/IR/1000154589. URL: https://publikationen.bibliothek.kit.edu/1000154589 (visited on 11/07/2023).

[19] Mathias Ziebarth. “Wahrnehmungsgrenzen kleiner Verformungen auf spiegelnden Oberflächen”. de. In: (2019). Medium: PDF Publisher: KIT Scientific Publishing. DOI: 10.5445/KSP/1000090271. URL: https://publikationen.bibliothek.kit.edu/1000090271 (visited on 11/07/2023).

Automated Security Analysis for Industrial Control Systems based on MITRE ATT&CK and IEC 62443

Jonas Vogl

KASTEL
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
Jonas.Vogl@kit.edu

Abstract

In this article a lightweight approach to automatically analyze the architecture of Industrial Control Systems (ICS) for cybersecurity issues is presented. The goal is to support network architects and administrators with identifying security weaknesses in their network architecture and help them find efficient solutions. For this a mapping between the attacker focused MITRE ATT&CK Framework [9] and the defense oriented IEC 62443 standard [5] is created. This mapping is then used to estimate for which attack techniques defenses are already in place or have to be improved.

1 Introduction

To set up an effective and efficient defense, it is vital to have good information on the assets and systems that have to be defended as well as the attacker. This is no new concept to the cybersecurity community and there are numerous projects that collect security relevant data. For this article mainly two of those projects are relevant. The MITRE ATT&CK Framework [9] is a collection of empirical data of attack behavior, which is used to model attackers in this work. The IEC

62443 standard [5] contains a collection of mitigations and recommendations when to employ them. IEC 62443 is used here to model the system under consideration (SUC). Besides that there are the Common Vulnerabilities and Exposures (CVE) [2] and Common Weakness Enumeration (CWE) [3] databases and several security standards that collect security best practices.

While there are many sources of information about many different aspects of cyber security, they tend to have their own focus. Many can be categorized as either attacker focused (CVE, CWE, MITRE ATT&CK) or defender focused (standards such as 62443, NIST CSF[11]). Combining these different sources is potentially valuable for security, but also tedious. There are several different works from recent years that map different databases frameworks and standards [7, 4, 8].

Another observation about especially the attacker centric databases is that they focus on finding and collecting vulnerabilities. Finding and closing vulnerabilities is an essential part of cyber security. However it forces the defender into a reactionary role where the defender waits for new vulnerabilities to be found and then close them. More proactive methods that can help improve a systems security without relying on knowledge of new vulnerabilities would be a good addition to defensive tools available. Other work in this direction is for example MITRE's Infrastructure Susceptibility Analysis and Assessment [10] project which is an attempt to forecast attacker behavior. It is in an early stage with no published results yet though.

This article contributes to the ongoing effort to better link and use the different existing knowledge bases by proposing a method to automatically analyze a system's security without relying on known vulnerabilities. A mapping of mitigations standardized in 62443 to the attack techniques collected in the MITRE ATT&CK framework is used to find weaknesses in a system's security concept based on the security requirements found in that system's risk assessment.

1.1 IEC 62443

IEC 62443 Security for Industrial Automation and Control Systems [5] is a collection of standards that codifies best practice knowledge specifically for

cybersecurity of ICS. While it's different parts are intended for different actors like device manufacturer, system owner/operator or integrator, IEC 62443's focus is on defense. Depending on the level of security required 62443 specifies mitigations that have to be implemented. While 62443 specifies what has to be done, it does not specify how. That means 62443's requirements can be met not only by technical measures, physical measures or policies such as fences or restrictions on the use of private devices are equally valid as well.

For this work we focus on 62443-3-3 of the standard, which describes security measures that are used to protect systems [6]. It is targeted towards operators and integrators of ICS and specifies which mitigations have to be implemented to achieve a given Security Level.

Security Levels in 62443. In 62443-3-3 mitigations are tied to Security Levels (SL). As the Security Level increases, 62443-3-3 specifies additional mitigations to defend against stronger attackers, that are associated with the higher Security Level. An overview of the 4 different security levels in 62443 can be found in Table 1.1. The main characteristic of each SL that differentiates it from the other SLs is highlighted in italic. SL 1, the lowest security level protects only against random events such as random bitflips that can happen on any carrier or natural disasters. So SL 1 is mostly concerned with safety, not security. From SL 2 on intelligent attackers of low motivation and skill are considered. Starting with SL 3 it is assumed attackers have ICS specific knowledge, which allows for more sophisticated attacks. On the highest Security Level, SL 4, Advanced Persistent Threats (APT) such as criminal organisations or government agencies are considered. They have significantly higher motivation and resources compared to other attackers.

Requirements in 62443-3-3. Part 3-3 of IEC 62443 specifies the requirements that have to be implemented in an ICS network to achieve a given security level. These are mitigations, organized in Foundational Requirements (FR), System Requirements (SR) and Requirement Enhancements (RE).

Foundational Requirements (FR) are broader categories that group SRs together based on the security goal they are supposed to achieve. Examples for FRs are Use Control, System Integrity or Restricted Data flow.

Security Level	Threat Type	Attacker Motivation	Attacker Skill
SL 1	*Unintentional*	None	None
SL 2	*Intentional*	Low	Low
SL 3	Intentional	Moderate	*ICS Specific*
SL 4	Intentional	*High*	ICS Specific

Table 1.1: Security Levels in IEC 62443

System Requirements (SR) are the base mitigations that can be implemented in an ICS network. Each SR is assigned a Security Level, usually SL 1 or 2. To achieve a Security Level all assigned SRs have to be implemented. Examples of SRs are authorization enforcement for all human users, auditable events or session locks. All three examples are assigned to FR 2 Use Control.

Requirement Enhancements (RE) are additional mitigations that improve on a given SR. They are also assigned Security Levels. Usually they are required to increase the SL of a SR to higher SLs of 3 or 4. As an example consider the enhacements to SR 2.1 authorization enforcement for human users, which is required from SL 1 on. To increase the SL to 2 the REs authorization enforcement for all users and a role that can assign permissions have to be implemented. To increase SL further supervisor override (SL3) and dual approval (SL4) need to be added.

Mitigations in 62443 are not all aimed at preventing an attack, but also at detecting an attack and recovering from it.

1.2 MITRE ATT&CK Framework

The MITRE ATT&CK framework [9]is a large knowledge base in which attacker behaviour is collected. The structure and design goals of the MITRE ATT&CK framework is explained in [1] [12].

The framework is structured into tactics, techniques and procedures. Tactics represent the different goals an attacker might have to achieve for a successful attack such as initial access, defense evasion, or impact. Not every attack has

to achieve all tactics. For each tactics there exist several different techniques attackers use to achieve the tactic. Some example techniques that belong to the impact tactic are Denial of View, Loss of Control or Damage to Property. While a tactic represents what an attacker wants to achieve, the techniques are how an attacker can achieve the tactic.

The third part of the MITRE ATT&CK structure is the procedure. Procedures represent the implementations of techniques. So for each technique there can be several procedures. Procedures are what is observed ”in the wild”, tactics and techniques are abstractions of that to better understand the attacker intent behind the procedure and to categorize them.

2 Automated Security Analysis

Here a concept for automated security analysis is presented. It consists of three parts, the system under consideration (SUC), the attacker model and the analysis method. The SUC is modeled in terms of mitigations that haven been deployed, as IEC 62443 describes them. The attacker model consists of a list of MITRE Techniques. The final component is the mapping of IEC 62443 and MITRE ATT&CK that ties attacker model and SUC together. All three parts are described in more detail below.

System Under Consideration. The SUC is modeled in terms of the mitigations that are implemented, as they are described in IEC 62443. All information required about the SUC for this analysis is a list of the SR/RE that are implemented in the SUC. While this SUC lacks technical details because of its abstraction, this also allows proactive scanning, without knowledge of vulnerabilities in the system. This also allows to factor in non-technical mitigations.

Attack Model. For the attacker model we use MITRE ATT&CK Techniques. Techniques are on a level of abstraction similar to that of 62443s SRs. An attack is represented by a list of all techniques used in the attack.

Mapping MITRE techniques to 62443. To analyze the threat a set of MITRE Techniques poses to a SUC, a connection between mitigations and attack techniques is required. For this a mapping between MITRE Techniques and SR/REs

is created. Each pair of technique and mitigation (SR/RE) can have one of two types of relationships. A technique can either *circumvent* a mitigation or a mitigation *mitigates* a technique. For example the technique deobfuscation (T1140) **circumvents** the RE ”Malicious code protection on entry and exit nodes” by hiding malicious code in seemingly innocent files. On the other hand the deobfuscation technique can be **mitigated** by ”Zone boundary protection” (SR 5.2). It is also possible that a pair of technique and SR/RE does not directly influence each other. Note that this mapping does not represent guarantees that a circumvention or mitigation is always successful but rather that a circumvention or mitigation is likely.

The resulting mapping can be thought of as the rules for a game of Rock, Paper, Scissors, where a technique wins (circumventing them) against some SR/REs while loosing to others (being mitigated by them). This mapping can then be used to compare a given system against an attack (a list of techniques) to find out if mitigations against the attack are in place or if defenses are lacking.

2.1 Usage

This analysis relies on an already conducted risk assessment. Guidance on how to conduct a risk assessment can be found for example in IEC 62443-3-2 [5]. During a risk assessment, already deployed security measures are identified as well as relevant attacks. This information can then feed the security analysis proposed here. Relevant attacks then have to be described as collection of MITRE techniques. And if the risk assessment was not already done according to 62443 some additional work is necessary to translate mitigations to SR/REs. Overall the overhead of using this approach is low, since all information required is collected during a risk assessment process anyways.

By comparing the lists of implemented SR/REs against relevant techniques it is possible to estimate whether an attack is already reasonably mitigated or requires more mitigations against it. The mapping as well as the mitigations in MITRE associated with a technique can be used as guidance on how to improve security.

3 Outlook

In this article a method to proactively analyze security of ICS systems was presented. It uses relatively abstract information about both attacks and SUC. This makes the method easy and cheap to employ. But that also means that results are abstract and ignore some factors. This abstraction allows this method to be used proactively, without taking vulnerabilities into account. After this initial method is implemented and evaluated in a lab environment it will be enhanced by adding information and missing factors to the models, depending on the results of the evaluation. Some missing factors that can be added are more details about SUC, such as network architecture, devices and vulnerabilities. On the attacker side it might be useful to also take into account in which order different tactics have to be achieved for a successful attack.

Other recent works, such as those of Kuppa et.al. [7], Grigorescu et.al. [4] or Kwon et.al. [8] map MITRE to other cybersecurity resources such as the CVE database or the NIST Cybersecurity framework. MITRE's Infrastructure Susceptibility Analysis and Assessment [10] also attempts to improve to use security information predictively. So this work falls into a recent effort of different actors to improve the usage of security information leveraging MITRE.

References

[1] Blake E. Strom et al. *ATTACK Design and Philosophy March 2020 Revision*.

[2] CVE. *Common Vulnerabilities and Exposures*. URL: https://www.cve.org/.

[3] CWE. *Common Weakness Enumeration*. URL: https://cwe.mitre.org/.

[4] Octavian Grigorescu et al. "CVE2ATT&CK: BERT-Based Mapping of CVEs to MITRE ATT&CK Techniques". In: *Algorithms* 15.9 (2022). ISSN: 1999-4893. DOI: 10.3390/a15090314. URL: https://www.mdpi.com/1999-4893/15/9/314.

[5] ISA. *ISA-62443 Security for Industrial Automation and Control Systems*.

[6] ISA. *ISA-62443-3-3 Part 3-3: System security requirements and security levels*.

[7] Aditya Kuppa, Lamine Aouad, and Nhien-An Le-Khac. "Linking CVE's to MITRE ATT&CK Techniques". In: *Proceedings of the 16th International Conference on Availability, Reliability and Security*. New York, NY, USA: ACM, 8172021, pp. 1–12. ISBN: 9781450390514. DOI: 10.1145/3465481.3465758.

[8] Roger Kwon et al. "Cyber Threat Dictionary Using MITRE ATT&CK Matrix and NIST Cybersecurity Framework Mapping". In: *2020 Resilience Week (RWS)*. 2020, pp. 106–112. DOI: 10.1109/RWS50334.2020.9241271.

[9] MITRE. *MITRE ATT&CK Framework*. URL: https://attack.mitre.org/.

[10] MITRE. *MITRE Infrastructure Susceptibility Analysis and Assessment*. URL: https://www.mitre.org/news-insights/fact-sheet/infrastructure-susceptibility-analysis-and-assessments.

[11] NIST. *Framework for Improving Critical Infrastructure Cybersecurity*.

[12] Otis Alexander, Misha Belisle, and Jacob Steele. *ATTACK for ICS - Design and Philosophy*.

Incorporating Causal Prior Knowledge into Deep Neural Networks

Shahenda Youssef

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
shahenda.youssef@kit.edu

Abstract

Deep Neural Networks have achieved significant success in solving complex problems across various domains due to their ability to capture complicated patterns in large datasets; however, they often require large amounts of data to learn effectively and often lack transparency in their decision-making processes, relying heavily on correlation rather than causation. Such limitations have led to incorporating causal Prior Knowledge into neural network models which stands as a significant advancement in machine learning, such knowledge can mitigate this data dependency, guide the learning process, and enhance not only the robustness and generalizability of models but also their interpretability and explainability. Additionally, it enables models to adapt to new tasks and domains with greater ease and effectiveness.

This report tackles the importance of incorporating causal prior knowledge into deep neural networks and the methodologies that facilitate this incorporation. Fundamental concepts of causality are reviewed, with emphasis on its importance for advancing AI towards causal representation learning.

1 Introduction

Although Deep Neural Networks (DNNs) have shown promising results in diverse domains, they still present limitations in several aspects that are left to be resolved. The insufficient amount of training data usually hinders its performance due to the lack of generalization, and the black-box nature of deep neural networks does not allow for a precise explanation behind its mechanism, preventing a new scientific discovery. They can discover features hidden within input data together with their mutual co-occurrence. However, they are weak at discovering and making explicit hidden causalities between the features. To overcome these challenges, It is critical to incorporate causality into DNNs framework [32].

The emergence of causality in AI signifies a paradigm shift from predictive to prescriptive analytics, where machines not only forecast but also recommend actions that lead to desired outcomes. Unlike correlation, which captures coincidental patterns, causality delineates a roadmap of cause and effect, empowering AI with the ability to reason beyond the data it is trained on [25].

Incorporating causal prior knowledge into the architecture of neural networks contributes to model robustness and generalizability, allowing models to better handle changes in data distributions by focusing on causal relationships rather than correlations [31]. Embedding causal reasoning helps models provide interpretability and explanations that resonate with real-world causality, aligning more with human thinking, and extending their use to interventions and policy-making [32].

The methodologies to incorporate prior knowledge within neural networks are varied [8]. Designing network architectures that detect complex data patterns. Imposing informed constraints on the loss function directs the optimization process towards solutions that respect established relationships and theoretical frameworks. Employing data augmentation and leveraging the insights from transfer learning further exploit the breadth of existing data and pre-trained models, accelerating the learning process. Knowledge graphs are adopted to enhance neural networks with information about relations between instances [1].

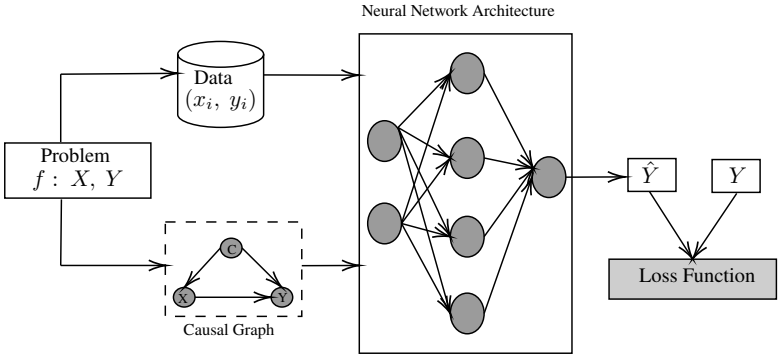


Figure 1.1: The proposed framework to incorporate causal prior knowledge into neural network.

These strategies encapsulate a concerted move towards data-driven learning with causal prior knowledge.

Current research focuses on enhancing model performance and explainability through the incorporating of prior knowledge into the learning process [36, 2, 10]. However, the development of models that incorporate causal prior knowledge continues to be a challenge. Figure 1.1 is a proposed framework to incorporate causal prior knowledge into a neural network, such a collaboration system can be achieved by involving the usual training data and additional prior knowledge that comes from an independent source, which is given by the causal graph model.

The rest of the report is organized as follows: Section 2 provides an overview of causality. Section 3 describes incorporation of prior knowledge. Section 4 addresses the related work in causal representation leaarning. Section 5 outlines some major challenges related to the incorporation of Causal prior knowledge into DNNs. In this section, we also present insights into potential directions for future research.

2 Causality

The study of causality seeks to establish the nature and strength of cause-and-effect relationships [26]. "Correlation does not imply causation" [25], two variables y and x could be correlated (statistically dependent) and, therefore, seeing x allows predicting the value of y, but if y is not caused by x then setting the value of x won't affect the distribution of y.

Causal inference is the process of concluding a causal connection based on the conditions of the occurrence of an effect. It involves establishing that a change in one variable (the cause) brings about a change in another variable (the effect). Researchers have developed methodologies to estimate causal effects from observational data. In this section, we present our interest frameworks that are introduced to causal inference.

2.1 Structural Causal Model

Structural Causal Models (SCM) provide a mathematical framework to model and infer causal relationships [24]. They are based on the idea that causal relationships can be represented by a set of structural equations and Directed Acyclic Graphs (DAGs). Each node in the DAG represents a variable, and each edge represents a causal influence from one variable to another. SCM can be represented as

$$X := f_X(PA_X, U_X), \tag{2.1}$$

a variable X in the causal graph is determined by a function f_X that could be linear or non-linear, whose inputs are its parents PA_X and a random variable U_X representing potential chaos and variables unobserved in the causal graph explicitly.

2.2 Average Treatment Effect

The potential outcomes framework [30] is used to estimate the causal effect of an intervention. Consider a binary treatment variable T, where $T = 1$ if the treatment is given and $T = 0$ otherwise. For each individual i, there are

two potential outcomes: $Y_i(1)$ is the outcome if the individual i receives the treatment, and $Y_i(0)$ is the outcome if they do not. The Individual Treatment Effect (ITE) for i would be

$$ITE_i = Y_i(1) - Y_i(0) \tag{2.2}$$

However, we never observe both potential outcomes for the same unit. This problem is known as the "Fundamental problem of causal inference" [24]. Since we cannot observe both potential outcomes for the same unit, we often focus on the average effect of the treatment across all units. The Average Treatment Effect (ATE) is defined as

$$ATE = \mathbb{E}[Y|\text{do}(T = 1)] - \mathbb{E}[Y|\text{do}(T = 0)], \tag{2.3}$$

where $\mathbb{E}[Y|\text{do}(T = t)]$ represents the expectation of the outcome Y under the intervention $do(T = t)$, do-operator allow for the identification and estimation of causal effects from observational data under certain conditions.

2.3 Propensity Score Matching

Propensity Score Matching (PSM) is a statistical technique used to estimate the effect of a treatment, policy, or other intervention by accounting for the covariates that predict receiving the treatment. The key idea is to match units that received the treatment with similar units that did not receive the treatment based on their propensity scores. The propensity score for a unit is the probability of receiving the treatment given a set of observed covariates. First, the propensity score $e(X)$ for each unit is estimated, typically using logistic regression for binary treatments

$$e(X) = P(T = 1|X) = \frac{1}{1 + e^{-(\alpha+\beta X)}} \tag{2.4}$$

where T represents the treatment assignment, X represents the covariates, α is the intercept, β is the vector of coefficients, and e is the base of the natural logarithm. After estimating the propensity scores, units are matched. The goal

is to find for each treated unit i a control unit j such that their propensity scores are as close as possible

$$\min_{i,j} |e(X_i) - e(X_j)|, \tag{2.5}$$

where $e(X_i)$ is the propensity score of the treated unit i and $e(X_j)$ is the propensity score of the control unit j.

Estimation of Treatment Effect, the Average Treatment Effect on the Treated (ATT) is often estimated by comparing the outcomes Y between matched units

$$ATT = \frac{1}{N_T} \sum_{i \in T} (Y_i - Y_{j(i)}), \tag{2.6}$$

where N_T is the number of treated units, Y_i is the outcome for treated unit i, and $Y_{j(i)}$ is the outcome for the control unit j that is matched to i.

3 Prior Knowlege Incorporation

The incorporation of prior knowledge into the construction of deep neural networks (DNNs), focuses on the nature of input data to a deep neural network, the loss function employed during training, and the model architecture or its parameters of the neural network [8, 9]. Ongoing studies are concentrated on combining methods to guarantee that the embedded prior knowledge effectively guides the learning process while still permitting the neural network to discover new data-driven patterns.

3.1 Input Data

Embedding domain knowledge into DNNs by transforming the input data, we discuss two ways to do this. One way is feature engineering is a key approach, where additional attributes derived from physics-based models are integrated with the training data. The training data is processed through domain-specific functions for embedding prior knowledge into deep learning. Feature engineering was found to be one of the most common ways of integrating prior knowledge into deep learning [16].

The other way is how to represent domain knowledge that takes the form of graph-based data as input, Knowledge graphs can be directly utilized by specialized deep network models such as Graph Neural Networks (GNNs), which process graph-structured data. These networks aggregate and synthesize information from the knowledge graph to enhance predictive tasks [6, 13, 17].

3.2 Loss Function

DNNs can be enhanced with domain knowledge by adding penalty terms to the loss function that enforce constraints derived from that knowledge [22, 12]. There are two primary types of constraints: syntactic, which are often implemented through regularization to control model complexity, and semantic, which encode domain-specific truths and logic [9]. Syntactic constraints are implemented by incorporating regularization terms into the loss function to control model complexity, such as the number of layers or parameters. It also involves an embedding approach, which is a lower-dimensional representation of discrete variables. Penalty terms based on regularizing embeddings are used to encode syntactic constraints on the complexity of embeddings to define prior parameter distributions using knowledge graphs embeddings [33]. Semantic constraints are imposed by the domain knowledge and can specify the conditions that model predictions must satisfy, such as falling within a certain numerical range.

When learning a function f from data (x_i, y_i), where x_i are input features and y_i are the actual labels, the generic hybrid loss function of the deep learning model

$$\arg\min_{f} \text{Loss}(Y, \hat{Y}) + \lambda R(Y, \hat{Y}) + \lambda_D \text{Loss}_D(\hat{Y}), \tag{3.1}$$

where $\text{Loss}(Y, \hat{Y})$ is the label-based loss, $Y, \hat{Y}$ are the actual labels and predicted values, respectively. $\lambda R(Y, \hat{Y})$ is a regularization function used to control model complexity. $\text{Loss}_D(\hat{Y})$ is the prior knowledge directly incorporated into the NN loss function and is used to enforce the model to respect the prior knowledge

while training. λ_D is a hyper-parameter determining the effect of domain loss in the objective function.

3.3 Network Architecture

DNNs can be enhanced by incorporating domain knowledge, either through constraining model parameters or by deliberate architectural choices.

Priors can be introduced on the parameters of a network. Explicitly, these would take the form of a prior distribution over the values of the weights in the network. The priors on networks and network weights represent our expectations about networks before receiving any data and correspond to penalty terms or regularizes. Two main methods have been used in DNNs, Transfer learning and Data Augmentation. Transfer learning is a technique to import weight priors in scenarios where data is scarce for the target problem. This method leverages existing models from a related source problem to inform the target model's structure or parameters, thus embedding the domain knowledge into the target domain [20, 27].

Data Augmentation based on prior information can effectively extend the original dataset with synthetic or transformed examples that reflect domain-specific insights. This method allows the integration of additional contextual or structural information, informed by prior understanding, to enrich the training process and improve the robustness and generalization of the neural network models [3].

Further, specialized structures in DNNs are enhanced when the architecture of the network is informed by domain knowledge, as the way knowledge representations are integrated into the network is largely determined by the type of architecture [16] such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Neural Networks (GNNs).

4 Causal Representation Learning

The work by Deng et al. [11] introduces a deep learning framework for societal event forecasting that leverages causal inference, and employs Individual Treat-

ment Effects (ITE) to estimate the influence of various treatments (or events) on societal outcomes with spatiotemporal environments. The model predicts potential outcomes for various treatment scenarios, thus incorporating causal information into event predictions. This process consists of two methods:

Approximation constraints l, u apply to event prediction scores $\hat{y}$, ensuring that the training of event predictors P for a location M at time $t+\delta$ takes into account the potential outcomes estimated by the causal inference model

$$l^{t+\delta} = \min(\hat{y}^{t+\delta}), \quad u^{t+\delta} = \max(\hat{y}^{t+\delta}). \tag{4.1}$$

The constraints limit the range of the ITE by enforcing minimum and maximum values derived from causal knowledge, where $\hat{y}^{t+\delta}$ is a set of potential outcomes for all treatment events, and the defined constraint loss term

$$\mathcal{L}_{\text{CSTR}} = \sum_{t \in T} \sum_{i \in M} \text{ReLU}(l_i^{t+\delta} - \hat{y}_i^{t+\delta}) + \text{ReLU}(\hat{y}_i^{t+\delta} - u_i^{t+\delta}). \tag{4.2}$$

By minimizing the total loss while training the predictor

$$\mathcal{L}_{\text{EVT}} = \mathcal{L}_{\text{PRED}} + \mu \cdot \mathcal{L}_{\text{CSTR}}, \tag{4.3}$$

where $\mathcal{L}_{\text{PRED}}$ is the loss function defined by the predictor P and μ is a hyperparameter.

Feature reweighting involves using the ITE estimated from the causal inference model to reweight event frequency features. This reweighting is essential for capturing the importance of features for predicting events. The approach defines a gating feature $\rho^{t+\delta}$ based on ITE, where for the j-th treatment event, the estimated ITE of a location at time $t+\delta$ is computed as follows

$$\hat{\tau}_{(j)}^{t+\delta} = \hat{y}_{(j)}^{t+\delta}(1) - \hat{y}_{(j)}^{t+\delta}(0), \tag{4.4}$$

where $\hat{y}_{(j)}^{t+\delta}(1)$, $\hat{y}_{(j)}^{t+\delta}(0)$ are the predicted potential outcomes with and without the treatment, respectively. The gating variables are applied to the original event frequency features through a sigmoid function σ to obtain a soft gated signal

$$\rho^{t+\delta} = \sigma(f_\tau(\hat{\tau}^{t+\delta})). \tag{4.5}$$

The event frequency features x are reweighted using the causal feature gates. The new feature vector is the element-wise product of the original feature vector x^t and the gating variables $\rho^{t+\delta}$

$$\tilde{x}^t = x^t \odot \rho^{t+\delta} + x^t \tag{4.6}$$

Such features are fed into a predictor to perform event prediction. They conducted extensive experiments on several real-world event datasets and showed that their approach achieves the best results in ITE estimation and robust event prediction involving multiple treatments and outcomes, which is considered an advancement over traditional correlation-based forecasting methods.

It is essential to recognize the expanding number of research that highlights the integration of causal regularization strategies into the framework of predictive modeling [19, 14, 15].

The paper by Teshima [35] introduces a model-independent method for data augmentation that leverages the conditional independencies relations in the data distribution encoded in causal graphs to enhance supervised learning.

Richens et al. [29] introduces the concept of counterfactual diagnosis, which uses counterfactual reasoning to evaluate the likelihood of a disease-causing the patient's symptoms. Structural causal models (SCMs) are discussed as methods for encoding the relationships between diseases, symptoms, and risk factors for more accurate diagnostic reasoning. The authors show that incorporating knowledge into machine learning can be effective in assisting medical diagnosis to reduce diagnostic errors.

Kyono et al. [18] demonstrates the utilization of causal graphs as prior knowledge to enhance model selection to enhance the robustness of neural network performance. By embedding this knowledge within a Structural Causal Model, derive a score that assesses the compatibility of a model's predictions with the SCM and input variables.

A recent work by Terziyan and Vitko [34] presents an approach for enhancing Convolutional Neural Networks (CNNs) by incorporating causality-awareness into the architecture. The authors introduce an architecture that includes an additional layer of neurons that is engineered to estimate asymmetric causality

in images using causal disposition [21] by using convolutional layers to capture features from images and then using these features to estimate conditional probabilities of the presence of one feature given another to improve image classification and generation. The causality map which is the calculated causality estimates is integrated into the CNN architecture, and the content of this map is calculated using

$$P(F^i|F^j) = \frac{\left(\max_{l,r=1,n} F^i_{l,r}\right) \cdot \left(\max_{l,r=1,n} F^j_{l,r}\right)}{\sum_{l,r=1}^{n} F^j_{l,r}} \tag{4.7}$$

where $P(F^i|F^j)$ is the causality map of size $k \times k$ (k - number of features), the features F^1, F^2, F^k represented by $n \times n$ feature maps. The causality map provides additional inputs to the network, which are used during backpropagation, enabling the network to discover which features have significant causal relationships. They also use it as a component within Generative Adversarial Networks (GANs) to enable the generation of images with respect to causalities. They demonstrated that their suggested model not only enhances the classification effectiveness of traditional CNNs but also improves the interpretability of the model's results.

5 Challenges and Future Prospects

Current research tends to address data-driven that is independent and identically distributed (IID). However, when dealing with spatiotemporal data that does not follow this IID assumption, the task of incorporating causal models that cope with strongly correlated values over time is not trivial [7].

The utilization of deep learning within manufacturing systems remains at an early stage, not only because of its solely data-driven nature, but also due to the limited research conducted on embedded causal knowledge into deep learning models by domain experts in the field.

Figure 1.1 raises some open research questions: incorporating causal prior knowledge requires modifications to the loss function, formulating an appropriate

term for the loss function can be complex. Introducing such a term frequently leads to complex optimization problems [5, 4].

New combinations of approaches are possible which have not been investigated yet, for example, by merging the causal prior knowledge with the DNNs architecture using the attention mechanism, the model iteratively processes the knowledge by selecting the relevant content at each step. The knowledge-based attention layer helps improve the prediction and the performance of the model.

Another prospective framework is to incorporate a causal graph model with an embedding graph layer, which would then serve as input to Bayesian Neural Networks (BNNs). This integration aims to enhance causal prior knowledge by refining the prior distribution during model training. This probabilistic approach reflects uncertainty in the model's predictions, where understanding the confidence level of a prediction is as important as the prediction itself [23, 28].

References

[1] Peter Battaglia et al. "Interaction networks for learning about objects, relations and physics". In: *Advances in neural information processing systems* 29 (2016).

[2] Katharina Beckh et al. "Explainable machine learning with prior knowledge: an overview". In: *arXiv preprint arXiv:2105.10172* (2021).

[3] Andrea Borghesi, Federico Baldo, and Michela Milano. "Improving deep learning models via constraint-based domain knowledge: a brief survey". In: *arXiv preprint arXiv:2005.10691* (2020).

[4] Luiz Chamon and Alejandro Ribeiro. "Probably approximately correct constrained learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16722–16735.

[5] Luiz FO Chamon et al. "Constrained learning with non-convex losses". In: *IEEE Transactions on Information Theory* 69.3 (2022), pp. 1739–1760.

[6] Qibin Chen et al. "Towards knowledge-based recommender dialog system". In: *arXiv preprint arXiv:1908.05391* (2019).

[7] Ricky TQ Chen et al. “Neural ordinary differential equations”. In: *Advances in neural information processing systems* 31 (2018).

[8] Tirtharaj Dash et al. “A review of some techniques for inclusion of domain-knowledge into deep neural networks”. In: *Scientific Reports* 12.1 (2022), p. 1040.

[9] Tirtharaj Dash et al. “How to tell deep neural networks what we know”. In: *CoRR, abs/2107.10295* (2021).

[10] Arka Daw et al. “Physics-guided neural networks (pgnn): An application in lake temperature modeling”. In: *Knowledge Guided Machine Learning*. Chapman and Hall/CRC, 2022, pp. 353–372.

[11] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. “Causal Knowledge Guided Societal Event Forecasting”. In: *arXiv preprint arXiv:2112.05695* (2021).

[12] Ethan Gallup, Tyler Gallup, and Kody Powell. “Physics-guided neural networks with engineering domain knowledge for hybrid process modeling”. In: *Computers & Chemical Engineering* 170 (2023), p. 108111.

[13] Manas Gaur, Keyur Faldu, and Amit Sheth. “Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?” In: *IEEE Internet Computing* 25.1 (2021), pp. 51–59.

[14] Dominik Janzing. “Causal regularization”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[15] Lucas Kania and Ernst Wit. “Causal Regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees”. In: *arXiv preprint arXiv:2205.01593* (2022).

[16] Sung Wook Kim et al. “Knowledge Integration into deep learning in dynamical systems: an overview and taxonomy”. In: *Journal of Mechanical Science and Technology* 35 (2021), pp. 1331–1342.

[17] Ugur Kursuncu, Manas Gaur, and Amit Sheth. “Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning”. In: *arXiv preprint arXiv:1912.00512* (2019).

[18] Trent Kyono and Mihaela van der Schaar. “Improving model robustness using causal knowledge”. In: *arXiv preprint arXiv:1911.12441* (2019).

[19] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. “Castle: Regularization via auxiliary causal graph discovery”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1501–1512.

[20] Xuan Liu, Xiaoguang Wang, and Stan Matwin. “Improving the interpretability of deep neural networks with knowledge distillation”. In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2018, pp. 905–912.

[21] David Lopez-Paz et al. “Discovering causal signals in images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6979–6987.

[22] Nikhil Muralidhar et al. “Incorporating prior domain knowledge into deep neural networks”. In: *2018 IEEE international conference on big data (big data)*. IEEE. 2018, pp. 36–45.

[23] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[24] Brady Neal. “Introduction to causal inference”. In: *Course Lecture Notes* (2020).

[25] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[26] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[27] Maithra Raghu et al. “Transfusion: Understanding transfer learning for medical imaging”. In: *Advances in neural information processing systems* 32 (2019).

[28] Qihan Ren et al. “A Representation Bottleneck of Bayesian Neural Networks”. In: (2022).

[29] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. “Improving the accuracy of medical diagnosis with causal machine learning”. In: *Nature communications* 11.1 (2020), p. 3923.

[30] Donald B Rubin. “Causal inference using potential outcomes: Design, modeling, decisions”. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[31] Bernhard Schölkopf. “Causality for machine learning”. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 765–804.

[32] Bernhard Schölkopf et al. “Toward causal representation learning”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[33] Naoya Takeishi and Kosuke Akimoto. “Knowledge-Based Distant Regularization in Learning Probabilistic Models”. In: *CoRR* abs/1806.11332 (2018). arXiv: 1806.11332. URL: http://arxiv.org/abs/1806.11332.

[34] Vagan Terziyan and Oleksandra Vitko. “Causality-Aware Convolutional Neural Networks for Advanced Image Classification and Generation”. In: *Procedia Computer Science* 217 (2023), pp. 495–506.

[35] Takeshi Teshima and Masashi Sugiyama. “Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 86–96.

[36] Laura Von Rueden et al. “Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.1 (2021), pp. 614–633.

An Evaluation of Large Language Models for Procedural Action Anticipation

Zeyun Zhong

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
zeyun.zhong@kit.edu

Abstract

This study evaluates large language models (LLMs) for their effectiveness in long-term action anticipation. Traditional approaches primarily depend on representation learning from extensive video data to understand human activities, a process fraught with challenges due to the intricate nature and variability of these activities. A significant limitation of this method is the difficulty in obtaining effective video representations. Moreover, relying solely on video-based learning can restrict a model's ability to generalize in scenarios involving long-tail classes and out-of-distribution examples. In contrast, the zero-shot or few-shot capabilities of LLMs like ChatGPT offer a novel approach to tackle the complexity of long-term activity understanding without extensive training. We propose three prompting strategies: a plain prompt, a chain-of-thought-based prompt, and an in-context learning prompt. Our experiments on the procedural Breakfast dataset indicate that LLMs can deliver promising results without specific fine-tuning.

1 Introduction

Understanding human activities from video data presents significant challenges due to the inherent variability and complexity of these activities. Traditional methods, which rely heavily on learning representations from large-scale video datasets, face two key limitations. First, the intricacy of human activities makes it difficult to obtain comprehensive representations, especially for longer videos. Second, dependence on extensive datasets restricts the models' ability to generalize to less common, long-tail classes and unseen scenarios.

Recent research has begun to explore the use of large language models [16] (LLMs) to overcome these challenges. These models, equipped with billions of parameters, can utilize training data aggregated from vast, unlabeled text corpora, and have demonstrated exceptional few-shot and zero-shot performance across various tasks. Prior methods [19, 12] have used LLMs for egocentric action anticipation, typically integrating an action recognition model to supply the LLMs with historical action sequence. However, this integration complicates the LLMs' process, leading to less intuitive results.

In this study, we aim to utilize LLMs for long-term action anticipation, minimizing the dependency on action recognition models. We evaluate procedural activities such as breakfast preparation, relying on ground-truth action histories. Our approach departs from traditional methods that are reliant on extensive video data and are limited by the respective training data distributions, focusing instead on the procedural knowledge and generalization ability of LLMs. We design three prompting strategies based on chain-of-thought [18] and in-context learning [1]. These strategies enable LLMs to anticipate future actions by providing a sequence of past observed actions in discrete text. Our experimental results on the Breakfast dataset demonstrate the effectiveness of LLMs in understanding the human activities, showcasing the potential of LLMs in a new domain of activity prediction and understanding.

2 Related Work

Action Anticipation aims to predict future actions given a video clip of the past and present. Many approaches initially investigated different forms of action and activity anticipation from third person video [7, 5, 9]. Recently, along with development of multiple challenge benchmarks [2, 3, 15, 10], the first-person (egocentric) vision has also gained popularity. To accurately predict future actions, the summarization of temporal progression of past actions is essential. To model the past action progression, earlier methods mainly used RNN [5, 6] or TCN [11]-based architectures, which have been shown to be inferior to the recent Transformer-based approaches [8, 21, 9, 22]. Based on the predicted time horizon, action anticipation approaches can be broadly grouped into two categories [20]: short-term anticipation approaches [2, 3] and long-term anticipation approaches [5, 10]. While short-term approaches predict actions a few seconds into the future, long-term approaches aim to predict a sequence of future actions (with their durations) up to several minutes into the future.

Large language models [1, 17] have significantly influenced the natural language processing (NLP) field, exhibiting an impressive capacity to generalize across unseen tasks. With their extensive training data and large parameter size, LLMs have demonstrated the ability to learn from examples provided in input prompts, a concept known as in-context learning [1]. Additionally, LLMs utilize a chain-of-thought [18] reasoning approach. This involves breaking down complex questions into simpler sub-questions, which are then sequentially addressed. This step-by-step reasoning enhances the accuracy and coherence of responses, especially for complex queries, and provides a transparent rationale for the model's thought process.

3 Method

To assess the effectiveness of LLMs in action anticipation, we utilize a procedural dataset, the Breakfast [13] dataset. The LLM is tasked with predicting future actions based on an input sequence of observed human actions, $[a_1, \ldots, a_M]$, where M represents the total number of observed actions. The objective is

Role:

```
{'role': 'system',
'content': 'You are a predictive AI assistant
    focused on Breakfast preparation. All fine-
    grained action classes are: [...].'}
```

Plain Prompt:

```
{'role': 'user',
'content': 'Given the observed fine-grained actions
    : [...], predict the next {N} actions using only
     the predefined action classes. Do not include
    actions outside the predefined list. Respond
    only in this format: <action1>, <action2>, ...'}
```

Figure 3.1: Plain setup. The LLM model is asked to predict N future actions based on a sequence of observations.

to forecast a subsequent series of N actions, $[a_{M+1}, \ldots, a_{M+N}]$, which the human actor is likely to perform. To achieve this goal, we introduce three prompts that are described in the following paragraphs. Additionally, we outline a post-processing methedology to align the outputs of LLMs with the desired format requirements.

Prompt Design. In our initial approach, we present a straightforward setup, as illustrated in Fig. 3.1. To improve the output quality of the LLM model (ChatGPT in our case), we configure the model as a predictive AI assistant specifically tailored for breakfast preparation tasks. To restrict the scope of predictions, we incorporate all action classes from the dataset into the model's setup. This prompt includes the task description and defines the input, i.e., a sequence of observations. In addition, the prompt also defines the output format, mandating the model to predict only actions that are contained in the predefined list.

In our second approach, we adopt a top-down approach [19], utilizing chain-of-thoughts prompts [18] (CoT), as illustrated in Fig. 3.2. This approach initially deduces the overarching activity from the history of actions and then formulates

Role:

```
{'role': 'system',
'content': 'You are a predictive AI assistant
    focused on Breakfast preparation. All fine-
    grained action classes are: [...]. All
    activities are: [...].'}
```

Intention Prompt:

```
{'role': 'user',
'content': 'Given the observed fine-grained actions
    : [...], identify the current activity based on
    these actions and then predict the subsequent {N
    } actions. Use only the predefined action and
    activity classes. Do not include actions or
    activities outside the predefined list. Respond
    only in this format: <activity>; <action1>, <
    action2>, ...'}
```

Figure 3.2: Top-down setup. The LLM model is first asked to identify the current activity given a sequence of observations, and then to predict N future actions based on both the inferred high-level activity and observations.

a plan considering both the historical actions and the intended goal. We construct two CoT questions: `Q1.What's the current activity according to previous actions?` `Q2.What are the future actions based on the inferred activity and previous actions?` To limit the predictive range for activity forecasts, we also incorporate all high-level activity classes into the model's setup.

In our last approach, we incorporate a few examples from the training set into the prompt to enable in-context learning [1] (ICL), as outlined in Fig. 3.3. Unlike fine-tuning, which involves backward passes through the entire or partial model, ICL leverages the inherent generalization capabilities of LLMs without being constrained to particular datasets or scenarios.

Inference of LLM and Post-processing. It is important to recognize that the outputs generated by LLMs may not always adhere to the required for-

```
'Given the observed fine-grained actions: [...],
    identify the current activity based on these
    actions and then predict the subsequent {N}
    actions. Use only the predefined action and
    activity classes. Do not include actions or
    activities outside the predefined list.
Example 1 - Observed: [...], Activity: [...],
    Predicted actions: [...].
...
Example 4 - Observed: [...], Activity: [...],
    Predicted actions: [...].
Respond only in this format: <activity>; <action1>,
     <action2>, ...'
```

Figure 3.3: Top-down prompt with in-context learning (ICL). A few examples are added to the prompt in Fig. 3.2 to enable in-context learning.

mat and taxonomy, even when the input prompts explicitly include classes from a predefined domain and request predictions within a certain format. For instance, the LLM model might predict `Activity: making tea\nNext predicted action: pour_water`, which deviates from the expected format of `<activity>; <actions>`. This discrepancy complicates the process of metric calculation. To address this, we implement a string matching rule to identify relevant activity or actions for metric evaluation. For simplicity, predictions that fall outside the predefined list are considered false predictions.

4 Experiments

Dataset. The Breakfast [13] dataset comprises 1,712 videos of 52 different individuals making breakfast in 18 different kitchens, totalling 77 hours. Every video is categorized into one of the 10 activities related to breakfast preparation. The videos are annotated by 48 fine-grained actions.

Prompt	Recognition	Anticipation		
	Top-1 ↑	Top-1 ↑	Top-1 agnostic ↑	Edit ↓
Plain	–	12.08±0.90	31.78±0.97	0.87±0.02
Top-down	62.01±2.37	14.83±0.93	33.41±0.90	0.84±0.01
Top-down ICL	94.07±1.14	35.05±1.20	66.46±1.50	0.55±0.02

Table 4.1: Comparison of three presented prompting strategies on the Breakfast [13] dataset. We report the mean performance and the standard deviation of five runs.

Metrics. We evaluate three metrics in this work: top-1 accuracy, top-1 order-agnostic accuracy, and edit distance (ED). Top-1 accuracy measures the precision of predictions in their chronological sequence. In contrast, top-1 order-agnostic accuracy focuses on the presence of correct predictions, regardless of their temporal order, reflecting scenarios where identifying future key actions is crucial, irrespective of preceding or succeeding actions. For instance, in robotic applications, foreseeing a need for assistance, such as an object hand-over, is pivotal, while the exact sequence of preceding or subsequent human actions is less critical. Additionally, we adopt ED [4, 14] to assess the sequential alignment of predictions with actual events. ED incorporates insertions, deletions, substitutions, and transpositions in the predicted actions. A lower ED indicates a higher similarity between the predicted and actual sequences.

Evaluation Details. In our experiments, we utilize the ChatGPT-3.5-turbo [16] model, to serve as the LLM model. The LLM model processes two observed past actions ($M = 2$) and forecasts the next N actions. N is a variable number in our setup and is set as the number of ground-truth actions minus the observed actions for each sequence. In the in-context learning (ICL) setup, illustrated in Fig. 3.3, we select four training examples, following [12]. Specifically, for each sequence in the test set, we identify diverse examples in the training set that include the observed test actions. If the number of identified training examples exceed four, only the initial four are chosen. Conversely, if less than four examples are found, all available examples are utilized in the ICL setup.

Results. In each experimental setup, we execute the Large Language Model (LLM) five times and present both the mean performance and the standard deviation in Table 4.1. Beyond evaluating the accuracy of future action predictions, we also compute the Top-1 accuracy for activity inference in top-down approaches. The results indicate a consistent enhancement in anticipation capabilities through the top-down method, which prioritizes identifying the current activity before predicting future actions. Furthermore, when in-Context learning (ICL) is activated using select examples from the training set, there is a notable improvement in both anticipation and activity recognition performance. Specifically, anticipation accuracy approximately doubles, and activity recognition accuracy increases by 52% (62.01 $\rightarrow$ 94.07). Additionally, the relatively narrow standard deviation across all metrics suggests that the LLM effectively leverages the provided context to refine its outputs.

5 Conclusion

In this study, we conduct an extensive evaluation of large language models, such as ChatGPT-3.5-turbo, focusing on their capability for long-term action anticipation, particularly leveraging their impressive zero-shot and few-shot learning abilities. This evaluation utilizes the procedural Breakfast dataset. Our findings indicate that these Large Language Models (LLMs) can accurately recognize current activities at an early stage and demonstrate commendable performance in predicting future actions. This underscores the potential of using LLMs for long-term anticipation tasks within the language domain. In the future, we aim to extend the application of LLMs to real-world anticipation scenarios by integrating an action recognition model.

References

[1] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[2] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *ECCV*. 2018, pp. 720–736.

[3] Dima Damen et al. “The epic-kitchens dataset: Collection, challenges and baselines”. In: *TPAMI* 43.11 (2020), pp. 4125–4141.

[4] Fred J Damerau. “A technique for computer detection and correction of spelling errors”. In: *Communications of the ACM* 7.3 (1964), pp. 171–176.

[5] Yazan Abu Farha, Alexander Richard, and Juergen Gall. “When Will You Do What? - Anticipating Temporal Occurrences of Activities”. In: *CVPR*. 2018. arXiv: 1804.00892.

[6] Antonino Furnari and Giovanni Farinella. “What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention”. In: *ICCV*. 2019.

[7] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. “RED: Reinforced Encoder-Decoder Networks for Action Anticipation”. In: *BMVC*. 2017.

[8] Rohit Girdhar and Kristen Grauman. “Anticipative Video Transformer”. In: *ICCV*. 2021. arXiv: 2106.02036.

[9] Dayoung Gong et al. “Future Transformer for Long-term Action Anticipation”. In: *CVPR*. 2022. arXiv: 2205.14022 [cs].

[10] Kristen Grauman et al. “Ego4D: Around the World in 3,000 Hours of Egocentric Video”. In: *CVPR*. 2022. arXiv: 2110.07058. (Visited on 04/28/2022).

[11] Qiuhong Ke, Mario Fritz, and Bernt Schiele. “Time-Conditioned Action Anticipation in One Shot”. In: *CVPR*. June 2019.

[12] Sanghwan Kim et al. “LALM: Long-Term Action Anticipation with Language Models”. In: *arXiv preprint arXiv:2311.17944* (2023).

[13] Hilde Kuehne, Ali Arslan, and Thomas Serre. “The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities”. In: *CVPR*. 2014.

[14] Vladimir I Levenshtein et al. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[15] Yin Li, Miao Liu, and James M Rehg. “In the eye of beholder: Joint learning of gaze and actions in first person video”. In: *ECCV*. 2018, pp. 619–635.

[16] OpenAI. *Chatgpt: Optimizing language models for dialogue*. 2022.

[17] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[18] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[19] Qi Zhao et al. “AntGPT: Can Large Language Models Help Long-term Action Anticipation from Videos?” In: *arXiv preprint arXiv:2307.16368* (2023).

[20] Zeyun Zhong et al. “A Survey on Deep Learning Techniques for Action Anticipation”. In: *arXiv preprint arXiv:2309.17257* (2023).

[21] Zeyun Zhong et al. “Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation”. In: *WACV*. 2023.

[22] Zeyun Zhong et al. “DiffAnt: Diffusion Models for Action Anticipation”. In: *arXiv preprint arXiv:2311.15991* (2023).

Karlsruher Schriftenreihe zur Anthropomatik (ISSN 1863-6489)

Band 1 Jürgen Geisler
Leistung des Menschen am Bildschirmarbeitsplatz.
ISBN 3-86644-070-7

Band 2 Elisabeth Peinsipp-Byma
Leistungserhöhung durch Assistenz in interaktiven Systemen zur Szenenanalyse. 2007
ISBN 978-3-86644-149-1

Band 3 Jürgen Geisler, Jürgen Beyerer (Hrsg.)
Mensch-Maschine-Systeme.
ISBN 978-3-86644-457-7

Band 4 Jürgen Beyerer, Marco Huber (Hrsg.)
Proceedings of the 2009 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-86644-469-0

Band 5 Thomas Usländer
Service-oriented design of environmental information systems.
ISBN 978-3-86644-499-7

Band 6 Giulio Milighetti
Multisensorielle diskret-kontinuierliche Überwachung und Regelung humanoider Roboter.
ISBN 978-3-86644-568-0

Band 7 Jürgen Beyerer, Marco Huber (Hrsg.)
Proceedings of the 2010 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-86644-609-0

Band 8 Eduardo Monari
Dynamische Sensorselektion zur auftragsorientierten Objektverfolgung in Kameranetzwerken.
ISBN 978-3-86644-729-5

Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Band 9 Thomas Bader
Multimodale Interaktion in Multi-Display-Umgebungen.
ISBN 3-86644-760-8

Band 10 Christian Frese
Planung kooperativer Fahrmanöver für kognitive Automobile.
ISBN 978-3-86644-798-1

Band 11 Jürgen Beyerer, Alexey Pak (Hrsg.)
Proceedings of the 2011 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-86644-855-1

Band 12 Miriam Schleipen
Adaptivität und Interoperabilität von Manufacturing Execution Systemen (MES).
ISBN 978-3-86644-955-8

Band 13 Jürgen Beyerer, Alexey Pak (Hrsg.)
Proceedings of the 2012 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-86644-988-6

Band 14 Hauke-Hendrik Vagts
Privatheit und Datenschutz in der intelligenten Überwachung: Ein datenschutzgewährendes System, entworfen nach dem „Privacy by Design" Prinzip.
ISBN 978-3-7315-0041-4

Band 15 Christian Kühnert
Data-driven Methods for Fault Localization in Process Technology. 2013
ISBN 978-3-7315-0098-8

Band 16 Alexander Bauer
Probabilistische Szenenmodelle für die Luftbildauswertung.
ISBN 978-3-7315-0167-1

Band 17 Jürgen Beyerer, Alexey Pak (Hrsg.)
Proceedings of the 2013 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-0212-8

Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Band 18 Michael Teutsch
Moving Object Detection and Segmentation for Remote Aerial Video Surveillance.
ISBN 978-3-7315-0320-0

Band 19 Marco Huber
Nonlinear Gaussian Filtering: Theory, Algorithms, and Applications.
ISBN 978-3-7315-0338-5

Band 20 Jürgen Beyerer, Alexey Pak (Hrsg.)
Proceedings of the 2014 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-0401-6

Band 21 Todor Dimitrov
Permanente Optimierung dynamischer Probleme der Fertigungssteuerung unter Einbeziehung von Benutzerinteraktionen.
ISBN 978-3-7315-0426-9

Band 22 Benjamin Kühn
Interessengetriebene audiovisuelle Szenenexploration.
ISBN 978-3-7315-0457-3

Band 23 Yvonne Fischer
Wissensbasierte probabilistische Modellierung für die Situationsanalyse am Beispiel der maritimen Überwachung.
ISBN 978-3-7315-0460-3

Band 24 Jürgen Beyerer, Alexey Pak (Hrsg.)
Proceedings of the 2015 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-0519-8

Band 25 Pascal Birnstill
Privacy-Respecting Smart Video Surveillance Based on Usage Control Enforcement.
ISBN 978-3-7315-0538-9

Band 26 Philipp Woock
Umgebungskartenschätzung aus Sidescan-Sonardaten für ein autonomes Unterwasserfahrzeug.
ISBN 978-3-7315-0541-9

Band 27 Janko Petereit
Adaptive State × Time Lattices: A Contribution to Mobile Robot Motion Planning in Unstructured Dynamic Environments.
ISBN 978-3-7315-0580-8

Band 28 Erik Ludwig Krempel
Steigerung der Akzeptanz von intelligenter Videoüberwachung in öffentlichen Räumen.
ISBN 978-3-7315-0598-3

Band 29 Jürgen Moßgraber
Ein Rahmenwerk für die Architektur von Frühwarnsystemen. 2017
ISBN 978-3-7315-0638-6

Band 30 Andrey Belkin
World Modeling for Intelligent Autonomous Systems.
ISBN 978-3-7315-0641-6

Band 31 Chettapong Janya-Anurak
Framework for Analysis and Identification of Nonlinear Distributed Parameter Systems using Bayesian Uncertainty Quantification based on Generalized Polynomial Chaos.
ISBN 978-3-7315-0642-3

Band 32 David Münch
Begriffliche Situationsanalyse aus Videodaten bei unvollständiger und fehlerhafter Information.
ISBN 978-3-7315-0644-7

Band 33 Jürgen Beyerer, Alexey Pak (Eds.)
Proceedings of the 2016 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-0678-2

Band 34 Jürgen Beyerer, Alexey Pak and Miro Taphanel (Eds.)
Proceedings of the 2017 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-0779-6

Band 35 Michael Grinberg
Feature-Based Probabilistic Data Association for Video-Based Multi-Object Tracking.
ISBN 978-3-7315-0781-9

Band 36 Christian Herrmann
Video-to-Video Face Recognition for Low-Quality Surveillance Data.
ISBN 978-3-7315-0799-4

Band 37 Chengchao Qu
Facial Texture Super-Resolution by Fitting 3D Face Models.
ISBN 978-3-7315-0828-1

Band 38 Miriam Ruf
Geometrie und Topologie von Trajektorienoptimierung für vollautomatisches Fahren.
ISBN 978-3-7315-0832-8

Band 39 Angelika Zube
Bewegungsregelung mobiler Manipulatoren für die Mensch-Roboter-Interaktion mittels kartesischer modellprädiktiver Regelung.
ISBN 978-3-7315-0855-7

Band 40 Jürgen Beyerer and Miro Taphanel (Eds.)
Proceedings of the 2018 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-0936-3

Band 41 Marco Thomas Gewohn
Ein methodischer Beitrag zur hybriden Regelung der Produktionsqualität in der Fahrzeugmontage.
ISBN 978-3-7315-0893-9

Band 42 Tianyi Guan
Predictive energy-efficient motion trajectory optimization of electric vehicles.
ISBN 978-3-7315-0978-3

Band 43 Jürgen Metzler
Robuste Detektion, Verfolgung und Wiedererkennung von Personen in Videodaten mit niedriger Auflösung.
ISBN 978-3-7315-0968-4

Band 44 Sebastian Bullinger
Image-Based 3D Reconstruction of Dynamic Objects Using Instance-Aware Multibody Structure from Motion.
ISBN 978-3-7315-1012-3

Band 45 Jürgen Beyerer, Tim Zander (Eds.)
Proceedings of the 2019 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-1028-4

Band 46 Stefan Becker
Dynamic Switching State Systems for Visual Tracking.
ISBN 978-3-7315-1038-3

Band 47 Jennifer Sander
Ansätze zur lokalen Bayes'schen Fusion von Informationsbeiträgen heterogener Quellen.
ISBN 978-3-7315-1062-8

Band 48 Philipp Christoph Sebastian Bier
Umsetzung des datenschutzrechtlichen Auskunftsanspruchs auf Grundlage von Usage-Control und Data-Provenance-Technologien.
ISBN 978-3-7315-1082-6

Band 49 Thomas Emter
Integrierte Multi-Sensor-Fusion für die simultane Lokalisierung und Kartenerstellung für mobile Robotersysteme.
ISBN 978-3-7315-1074-1

Band 50 Patrick Dunau
Tracking von Menschen und menschlichen Zuständen.
ISBN 978-3-7315-1086-4

Band 51 Jürgen Beyerer, Tim Zander (Eds.)
Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-1091-8

Band 52 Lars Wilko Sommer
Deep Learning based Vehicle Detection in Aerial Imagery.
ISBN 978-3-7315-1113-7

Band 53 Jan Hendrik Hammer
Interaktionstechniken für mobile Augmented-Reality-Anwendungen basierend auf Blick- und Handbewegungen.
ISBN 978-3-7315-1169-4

Band 54 Jürgen Beyerer, Tim Zander (Eds.)
Proceedings of the 2021 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-1171-7

Band 55 Ronny Hug
Probabilistic Parametric Curves for Sequence Modeling.
ISBN 978-3-7315-1198-4

Band 56 Florian Patzer
Automatisierte, minimalinvasive Sicherheitsanalyse und Vorfallreaktion für industrielle Systeme.
ISBN 978-3-7315-1207-3

Band 57 Achim Christian Kuwertz
Adaptive Umweltmodellierung für kognitive Systeme in offener Welt durch dynamische Konzepte und quantitative Modellbewertung.
ISBN 978-3-7315-1219-6

Band 58 Julius Pfrommer
Distributed Planning for Self-Organizing Production Systems.
ISBN 978-3-7315-1253-0

Band 59 Ankush Meshram
Self-learning Anomaly Detection in Industrial Production.
ISBN 978-3-7315-1257-8

Band 60 Patrick Philipp
Über die Formalisierung und Analyse medizinischer Prozesse im Kontext von Expertenwissen und künstlicher Intelligenz.
ISBN 978-3-7315-1289-9

Band 61 Mathias Anneken
Anomaliedetektion in räumlich-zeitlichen Datensätzen.
ISBN 978-3-7315-1300-1

Band 62 Jürgen Beyerer, Tim Zander (Eds.)
Proceedings of the 2022 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-1304-9

Band 63 Fabian Dürr
Multimodal Panoptic Segmentation of 3D Point Clouds.
ISBN 978-3-7315-1314-8

Band 64 Jutta Hild
Nutzung von Blickbewegungen für die Mensch-Computer-Interaktion mit dynamischen Bildinhalten am Beispiel der Videobildauswertung.
ISBN 978-3-7315-1330-8

Band 65 Jürgen Beyerer, Tim Zander (Eds.)
Proceedings of the 2023 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory.
ISBN 978-3-7315-1351-3

Lehrstuhl für Interaktive Echtzeitsysteme
Karlsruher Institut für Technologie

Fraunhofer-Institut für Optronik, Systemtechnik
und Bildauswertung IOSB Karlsruhe

In 2023, the annual joint workshop of the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB) and the Vision and Fusion Laboratory (IES) of the Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT) was hosted at a Schwarzwaldhaus near Triberg. For a week from the 30th of July to the 4th of August the doctoral students of both institutions presented extensive reports on the status of their research and discussed topics ranging from computer vision, security to machine learning and large language models.

The results and ideas presented at the workshop are collected in this book in the form of detailed technical reports. This volume provides a comprehensive and up-to-date overview of the research program of the IES laboratory and the Fraunhofer IOSB.

ISSN 1863-6489 (Schriftenreihe)
ISSN 2510-7259 (Tagungsband)
ISBN 978-3-7315-1351-3