Karlsruher Schriften
zur Anthropomatik

Band 51

Jürgen Beyerer, Tim Zander (Eds.)

Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

KIT Scientific Publishing

Jürgen Beyerer, Tim Zander (Eds.)

Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

Karlsruher Schriften zur Anthropomatik
Band 51
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 2020 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 2021 – Gedruckt auf FSC-zertifiziertem Papier

ISSN 1863-6489 (Schriftenreihe)
ISSN 2510-7259 (Tagungsband)
ISBN 978-3-7315-1091-8
DOI 10.5445/KSP/1000130397

Preface

In 2020, 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 for the first time due to the pandemic at the IOSB in Karlsruhe where the strict infection protection regulations of the KIT could be enforced.

For a week from the 27th to the 31st of July the PhD students of the both institutions delivered extended reports on the status of their research and participated in heated discussions on topics ranging from computer vision and optical metrology to usage control and neural networks. 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 the research program of the IES Laboratory and the Fraunhofer IOSB.

The editors thank Arno Appenzeller, Paul Wagner and the other organizers 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 the suggestions of their colleagues.

Prof. Dr.-Ing. habil. Jürgen Beyerer
Dr. Tim Zander

Contents

Preface I
Jürgen Beyerer and Tim Zander

Privacy and Patient Involvement in e-Health 1
Arno Appenzeller

Characterization of Mueller matrices in retroreflex ellipsometry 19
Chia-Wei Chen

Data Annotation Process for Activity Recognition in Public Places ... 33
Mickael Cormier

3D Semantic Segmentation with Twin-Representation Networks 53
Fabian Duerr

Crowd-level Human Keypoint Tracking 67
Thomas Golda

DOE-based Multi-spot Confocal Interference Microscope 83
Zheng Li

Discovering Causal Structures in Technical Systems 91
Josephine Rehak

A Step Towards Explainable Person Re-identification Rankings 107
Andreas Specker

Multi-object Tracking in Drone Videos 123
Daniel Stadler

Classifying Usage Control and Data Provenance Architectures 135
Paul Georg Wagner

Learning Universal Representation for Multi-formats 3D Objects . . . 155
Chengzhi Wu

Privacy and Patient Involvement in e-Health Worldwide: An International Analysis

Arno Appenzeller

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

Abstract

Nearly every nation is actively working on an e-Health policy or already has one. Personal Health Records (PHR) are considered as one of the key elements in the digitisation of the health sector. While nearly every e-Health agenda mentions privacy and data protection, the real world implementation can differ. A recent example is the planned launch of the German "Elektronische Patientenakte" (ePA), which has only limited data control features for the patient in its first version. This paper gives an overview of how the e-Health policies of the G7 nations handle patient involvement and privacy for their PHR projects. With this analysis we show that privacy and patient involvement are crucial for the acceptance of such projects. Finally we propose a data sovereignty framework with guidelines for PHRs to give a user control over his data and establish trust in such systems through broad access, fine granular control, informed decision making, intuitive user experience and comprehensible transparency.

1 Introduction

e-Health applications and their general use is becoming more and more common and widely available. Nearly every developed country has a national e-Health

strategy to make use of the data created during a patient's treatment. There are enormous expected benefits from the broad availability of electronic health data. From better data availability for medical research, paperless hospitals that can easily send data to a patient's general practitioner to patient centred care where a patient is in the middle of the treatment and is in possession and control of all data. While those are just examples it is clear that the potential benefits make e-Health projects worth to be pursued for a long time period. When looking at the German national e-Health project, the "Elektronische Patientenakte (ePA)", which is set to launch in 2021, it can be easily seen that patient involvement and privacy controls are controversial topics. To speed up the launch of the ePA there are several limitations to access right management for the patient in the initial version.[1] A fine granular access management is promised at a later time but with no concrete dates yet. This lead to a lot of controversy and discussion from several parties. The German Federal Commissioner for Data Protection and Freedom of Information Kelber announced that this violates the General Data Protection Regulation (GDPR) of the EU and that they will pursue legal checks before the ePA launches.[2] It remains to be seen if this will weaken public trust in the nationwide project. This is only one of the examples that shows that data sovereignty is an important topic when it comes to digital personal health records. According to GDPR Article 9, personal medical data is considered as sensitive data and must not be processed by default. To make processing of personal health data possible (or the other sensitive data, that is stated in Art. 9) one of the exclusions of Article 9 Paragraph 1 a) to j) must be fulfilled. One of these exclusions is the explicit declaration of consent to the data processing by the affected person. While there are different approaches to the term of data sovereignty, there is still no clear definition what it means regarding the processing of medical data. The previous example for Germany and the EU regulation gives a glimpse of what is considered important in the EU in those terms. However nearly every country's e-Health strategy includes a project similar to the German ePA. Many European countries also define their strategy

[1] https://www.heise.de/newsticker/meldung/Elektronische-Patientenakte-Datenhoheit-kommt-spaeter-4427379.html [Accessed 25 November 2020; In German]

[2] https://www.heise.de/news/Datenschutzbeauftragter-kuendigt-Massnahmen-gegen-Patientendatenschutzgesetz-an-4873642.html [Accessed 25 November 2020; In German]

with GDPR in mind, but it is interesting how other nations for example the US handle the topic of patient involvement, control and privacy of personal medical data. It is also important to analyse if the e-Health programs even have a patient centred approach or if the focus is more on the broad availability of the data for doctors and research. This paper will give an in-depth look on the e-Health strategies of the G7 nations and other examples. Building on this analysis of the different countries and especially the role of patient involvement and privacy in those programs, we will define criteria for data sovereignty for personal medical data. The paper is structured as follows: At first we will define the term data sovereignty and what it can mean for patient involvement. After this we will give an overview of e-Health in the G7 nations and other good examples. With the provided analysis of our findings, we will improve our data sovereignty definition and discuss our results. We close this paper with a conclusion and an outlook to further research, that is necessary in this area.

2 Data Sovereignty for Patient Involvement

The term data sovereignty is nowadays mostly used not only for personal data, but also for the processing of data as an economic good. However for both use-cases a wide range of control and possibilities to intervene is mandatory. When looking at the patient's perspective of data sovereignty, it is necessary to have a look at what existing regulations enable for the affected person. In the introduction it was already mentioned that the GDPR requires consent to process medical data. There are also some requirements for privacy consent in the GDPR. First every consent must have a specific purpose. Art. 5 Par. 1 b) says that a purpose has to be unambiguous and that the data is not allowed to be processed for something that is not consistent with the declared purpose. Furthermore the data should be limited to what is necessary for this purpose to enable data minimisation. Another requirement made in Art. 4 Par. 11 is that a consent should be freely given and express the affected person's explicit agreement to the data processing. Besides this it should be always possible for a subject to withdrawal its consent. Also the Recital 32 indicates that an opt-out principle for personal data is not allowed. So any pre-ticked boxes or the assumption, that

remaining silent means confirmation are not lawful. Finally the Article 15 of GDPR requires that a subject should have the right to access its data to see if data was processed and for what purpose it was done. It remains to be noted that while GDPR is a European regulation, the execution between countries can be different. For example Belgium explicitly requires written consent for the processing of medical data. Furthermore Belgium also requires direct access for the patients to their health data, while Portugal limits access to physicians. In Germany there are also specific laws and regulations that have implications on the execution on GDPR such as the local hospital regulations or the laws like the Patientendaten-Schutz Gesetz (PDSG) mentioned in the introduction. Even with the common GDPR the EU remains fragmented, which makes a closer look necessary [7]. In terms of digital consent the level of granularity is always an important topic. The best case would be an arbitrary level of granularity for consent. This means a patient could choose any data he wants to share with any third party the patient wishes. In reality those approaches are often limited. The limitations can have several reasons like technical limitations or limitations of user interfaces. The German law defines the basic control over personal data as information self-determination. While this term comes from an age before personal health records, the basic principles like control and transparency of data usage remains important. Therefore we consider data sovereignty as information self-determination + $\boldsymbol{X}$. While this $\boldsymbol{X}$ seems arbitrary it must be defined interdisciplinary. There are legal, ethical and technical considerations that needs to be done. For example not every legal definition can be executed in the exact same way technically and not every technical possibility comes without ethical concerns. However the technical side of data sovereignty has to enable everything needed in terms of consent, with a rich digital consent management, and transparency with possibilities to automatically track usage of personal health data. Nevertheless as described before technical solutions alone are not sufficient, therefore this paper gives an overview of the state of the art and suggests technical solutions that can help to improve data sovereignty.

3 e-Health in the G7 Nations

In this section we provide an overview of the e-Health projects in the G7 nations and other good examples. The overview will be focused on privacy and security aspects of the different projects.

3.1 Methodology

For our research we choose a qualitative approach. We focused on the G7 nations since that is where we expected the most resources available in English language versions. In addition we decided to include a few other notable examples we found during our research. For our systematic research approach, we used the directory of e-Health policies by the World Health Organization (WHO), which was created with information of the states itself or through an online search and research in academic literature [16]. The directory entry served as starting point for a general overview. In addition the European Union has a similar overview for a few selected nations [9]. For a more focused view on privacy policies regarding e-Health, we did our own literature and online research.

3.2 Germany

As mentioned in the introduction Germany has a lot of ongoing e-Health projects, but the largest is the national ePA, which is created by a consortium of different parties called gematik.[3] Data protection and privacy is a core topic while developing the ePA. This policy has a legal foundation in the so called e-Health law, which requires the highest priority for data protection from a legal and technical standpoint [2]. The gematik approach to fulfil this requirement is with the policy that no data should be accessed through the *public* internet and that a secure tunnel is required anytime [11]. Nowadays with the rise of mobile applications this approach is not fully valid anymore, since smartphone apps, that give access for a patient to his data, are planned. Besides this another

[3] `https://www.gematik.de`

use-case is data usage when visiting a doctor. Access to this data requires two factors from each party. One factor is the so called “Gesundheitskarte”, which is a chip card that is also proof of health insurance. When a patient wants to access his data, he also needs to enter a PIN code. In this scenario the doctor also needs to prove his/her identity and confirm access with his medical id card, which is similar to the “Gesundheitskarte”. In addition every access to data of the personal health record is recorded in a log file. This helps the patient to trace the usage of his data. In terms of patient control the ePA tries to do a staged rollout of control mechanism for the initial launch. In the first version only basic access control is possible. This means that a patient can give a doctor or a different party full access to every data or no access. This lead to a lot of controversy as described in the introduction of this paper. In addition the final planned stage will still lack the option of full control since access rights can only be managed for several document types like a doctor’s letter, that includes more than one medical observation and not a fine granular level for every medical resource of a patient. In 2014 Dehling et al. did an evaluation of the ePA project and compared it to their approach of patient-centred health information technology service [5]. They described several requirements on how sensitive medical information should be handled with a focus on privacy and security. Their result was that ePA was lacking a lot of things like the possibility for anonymous data sharing, unlinkability and some things like confidentiality, access control and authorisation are only partly fulfilled. It remains to be seen how well the German ePA will be accepted with all its privacy controversy, when it is set to launch in 2021.

3.3 France

France has several long term national e-Health strategies [8]. It is part of a digitisation agenda from 2011 where e-Health is a fundamental part. Inside this e-Health agenda there are five key points and privacy is one of them: “Open access to health data, respectful of personal information privacy, to serve the steering of the health care system, as well as public health and research (open data)” [13]. 2011 was also the year of the introduction of France’s first personal health record project, the so called Dossier Médical Personel (DMP). The main

principle of the DMP is not to be a complete health record of a patient, but rather an exchange and information platform where physicians can include the information from a patient they consider necessary. It is also a more document based platform that includes files like physician reports based on international standards like the Clinical Document Architecture (CDA). The whole DMP is opt-in and will be created by the general practitioner (GP) if a patient gives consent. In terms of patient involvement the DMP has an interesting history. After the initial launch only doctors were able to access a patient's DMP. There was no explicit way for a patient to access the records without a doctor. This caused a very low adaption of the platform, which lead to a relaunch that was more patient centred and provided direct access for the patient. With this higher adoption and acceptance was noticeable[3]. This version also included granular access right management for the patient. First a patient needs to authorise every physician, so that he can have access. Then every document can have a certain status. The status can be open so that everyone that has access to the DMP can access it. Another status is hidden, where only the patient, the physician that created the document and the GP see the resource. Other doctors see that there is a hidden file. Lastly there is a confidential status, which can be used when there is a sensitive diagnosis that should be viewed for the first time in the presence of the corresponding doctor. In addition to these access rights a patient can upload own documents and see an access log for every document.

3.4 Italy

Italy currently has no nationwide personal health record project with a focus on direct patient involvement. However there are different e-Health projects to introduce a nationwide electronic health record (EHR). The e-Health law sees three main tasks for an EHR: improvement of treatment, research and evaluation of the care quality. While there is no direct patient involvement, patients can control if there is an EHR and what data will be stored there with their consent. Besides from exceptions for patient care or treatment, the patient is in the center of the consent decision [1]. It needs to be mentioned that this is a legal requirement and the patient has no technical way to control this yet. In the past this lead to cases where EHRs were created without consent. Nevertheless

the patient has a right to get access to the EHR and privacy is a core principle in laws and strategies for e-Health. From a technical perspective there is still more to do to introduce the EHR and to improve direct patient involvement [10].

3.5 United Kingdom

The UK with its centralised National Health Service (NHS) launched several EHR systems in the past. In 2012 the government released a strategy paper that described a ten year framework for health and care [6]. Two key points were that the patient should be in the center of care and that digitisation should give benefits in a broad spectrum. This starts with tools for patients to make digital appointments, receive digital prescriptions and self-assessment tools. Furthermore standardised data communication allows data to improve the quality of care and reduction of inconsistent or incomplete data for health care providers. The strategy paper also described privacy concerns as potential issues. The paper stated the following position regarding those concerns: "not sharing information has the potential to do more harm than sharing it". However there are no more details besides that data should be shared in a confidential and private way and that patient should control this process. One of the already launched projects is the Care.data program.[4] The project aims to store all data of GPs in a common centralised database. Patients can deny consent to participate, however data will then be store anonymised. The focus of Care.data is on the secondary usage of data. Hoeksma took an in depth look at the program and its implications on privacy [12]. There is paid access to the platform and it offers matching of data from different sites to enable longitudinal tracking of patients' progress. This was done by using a patient identifier, the NHS number, date of birth, sex, ethnicity and postcode. After the linkage those identifiers are replaced by a pseudonym. Overall Hoeksma criticized the platform for its lack of transparency for the patient. All in all the missing transparency and other issues lead to the shutdown of the project. Another project that is still in use is the Summary Care Record. This service is a personal health record platform similar to the system of Germany's ePA. It is only created with the patient's consent, which can have

[4] `https://www.england.nhs.uk/2013/10/care-data/`

different levels. It is possible to consent to the storage of all information or only to allow the storage of necessary medical data. Currently the patient has no independent access to the platform, but can request the stored data from his GP. There are also some privacy concerns related to the service as the foundation medConfidential stated when they analysed conformance with the GDPR. [5] In addition there is access for secondary use for third parties. This is a default and needs explicit opt-out by the affected person. Like before the opt-out options have different stages. Opt-out can be universal or the patient can opt-out for every secondary use except when the explicit purpose is to provide his own treatment or care. The third-party access also lead to some controversy when life insurance or credit-card companies gained full access through subject access requests, that should only provide them the relevant data for the purpose.[6]

3.6 United States of America

The USA with its federal states has a variety of federal and national privacy laws, that lead according to Dumortier et al. to problems in regard of the introduction of e-Health [7]. Additionally many privacy laws are rather old and from a time before the digitisation. The major regulation in terms of health data is the Health Insurance Portability and Accountability Act (HIPAA), which has general privacy rules for personal health data. One of the main principles is that a patient must give his consent before health data is processed or given to another party. One exclusion is the usage in the context of a treatment or to process payment. However the patient should always have a certain degree of control for sensitive data. HIPAA also regulates when consent is needed and how a consent has to look like. It gives patients the right to access their data and the right to correct data. There are special requirements how to use data for secondary usage and who is allowed to do it. The HIPAA regulation is created to avoid that personal health data gets into the hands of unauthorised people. Besides privacy regulation HIPAA also has requirements in terms of

[5] `https://medconfidential.org`

[6] https://www.telegraph.co.uk/news/nhs/10855450/Probe-into-claims-that-insurers-given-access-to-full-medical-records.html [Accessed 25 November 2020]

security of the data. It remains to be noted that federal law can lead to exclusions and exceptions of the HIPAA regulation. This makes exchange of health data between federal states difficult. Besides national projects and strategic plans to boost the digitisation of the digital health sector, there is an existing solution called Blue Button.[7] This EHR service is created by the Blue Button initiative and is supported by many health providers on a voluntary basis. If there is a website that provides the possibility to safely download the personal health data of patients a blue button appears and enables the possibility to do so. The format does not necessarily have to be machine readable, it also can be a PDF or a some other document. Rather than providing a personal health record platform like in other nations, the intention is to make analogue exchange of health data easier and enable a possibility for the patient to get access to his data. There is also a successor in development called Blue Button+ that plans to enable a digital exchange with explicit control of the affected patient. In addition to the initiatives of the government there are many projects from private companies like Apple Health or Microsoft Health Vault to create personal health records.

3.7 Japan

Japan has a long history of the introduction of electronic health record systems. It already started in 1998 with the development of formats for the EHR. Those fundamental approaches had considerations about security but obviously did not look at patients in control of their data. There are regional differences in terms of e-Health usage. For example there are hospitals that offer their own EHR where patients have access but there is no nationwide personal health record. In 2018 the next-generation medical infrastructure law was introduced [15]. After a look at the status quo in terms of e-Health, which showed that only the minority of hospitals use EHR systems, that most of digital patient data remains unused and that documentation of patients' EHR is mostly incomplete, the new law allowed access to anonymised patient data for secondary usage like research without explicit consent of a patient. Nevertheless the patient will be informed about the usage of his data and can intervene - no intervention means consent.

[7] `https://www.healthit.gov/topic/health-it-initiatives/blue-button`

Most of the system and decisions how to use data remain in the responsibility of the participating hospitals. They need to decide how to anonymnize the data and if and how the patient should be involved. A study from Morris et al. showed that most of Japans population have privacy concerns and favor a system where they have control [14].

3.8 Canada

Canada has a privacy regulation called Personal Information Protection and Electronic Document Act (PIPEDA) similar to the HIPAA of the US. The goal is to protect personal health data from commercial usage without consent [4]. Furthermore there is Canada Infoway, which develops e-Health solutions for Canada.[8] A key goal is that the patient is in the center of the treatment. However there is currently no nationwide general personal health record.

3.9 Others

Besides the G7 nations there are other good examples for national e-Health initiatives with patient involvement. One example is Australia which has a system called My Health Record for a personal health record.[9] Privacy is a very strong requirement there and is enforced by law and technical requirements. Another example is Estonia, where there is one central platform where every health data is stored. In terms of privacy this project follows the rule of GDPR.

4 Analysis

When looking at the e-Health policies of the G7 nations and the role of privacy and patient involvement it becomes clear that nearly every nation applies some kind of general regulations like the GDPR or more health specific ones like HIPAA. In addition to those general regulations many nations have specific

[8] https://www.infoway-inforoute.ca/en/
[9] https://www.myhealthrecord.gov.au

laws for e-Health. For example Germany has a whole series with the recent PDSG as example. One interesting observation is that European countries have different executions of GDPR in terms of e-Health. Another observation is that there is a gap between technical execution and legal requirements. A good example where the technical planning is behind the law is the staged launch of the German ePA in terms of access control. A general impression is that nearly every personal health record is in an early phase. The DMP in France is a good example what impact patient involvement has on acceptance. This should be considered as cautionary example about how a good revamp with the patient in mind can look like. When looking at projects from the UK the importance of privacy can be shown by looking at projects cancelled due to privacy issues. Unfortunately there is no project that meets our requirement for data sovereignty in terms of data control and an arbitrary level of granularity to do so. The lack of complete data control is also related to open questions in terms of secondary usage and data donations. This is a very sensitive topic since there needs to be a fair trade off between data usage for research, which potentially benefits the general public in general and protecting data of individuals. The depth of this controversy can be easily seen by the discussion whether such processes should be opt-in or opt-out. With the general overview and some failed examples our research shows strong indications that it is important to involve the patient in a transparent way instead of hiding such data usage from him.

In general it is shown that technical solutions alone can only be partial solutions. Another major topic is to intuitively present the technical possibilities to the user through a suitable user experience. This should help the patient to be in the position to control his data, but also understand the implications of the control. To achieve this we propose a data sovereignty framework for personal health records that gives user control over his data and establishes trust in the system with the following properties.

- **Broad access:**
  Independent access is mandatory for every aspect a user needs for data sovereignty. Access should be as easy as using a dedicated app on the patient's smartphone. Access exclusively in presence of a physician should be discouraged because it will not let the user have an independent look on his data.

- **Fine granular control:**
  A user should be able to control every access to every datapoint of his PHR. When a consent for data usage is requested the user should be able to decide which data is allowed to be used by whom. This also leads to requirements for the underlying data structure. Documents that combine different data of patients should be avoided or also offer fine granular control.

- **Informed decision making:**
  Any decision which is not based on information or deeper knowledge can not be an informed decision. This property has consent decisions in mind, which can be very complex and hard to understand for a patient. A sovereign patient should be able to know what he does at any time. This should be supported by recommendation systems that evaluate decisions based on the patients preference. In the concrete case of consent this can be done by considering the preferences of the user, the requested data and the purpose of the request.

- **Intuitive User Experience:**
  All possibilities are limited when a user does not understand them. This is a rather interdisciplinary challenge from an ethical, legal, technical and design standpoint. For example an access system where the user feels overwhelmed by the possibilities will not help him to be sovereign. The mentioned discipline must define the requirements, so that the user has a good experience.

- **Comprehensible transparency:**
  A patient should be able to reproduce every usage of his data. He should not only be able to see the first data usage but also what such first data request imply, for example a research project that gives data to a partner to process it. This should be supported from a technical perspective and encouraged by guidelines how to use the data and make it transparent for the user.

We propose that a PHR that wants to enable patient involvement and data sovereignty follows those guidelines and execute and extend them according to the project's need.

5 Conclusion & Outlook

This paper gives an overview of privacy and patient involvement in e-Health worldwide. We see patient involvement, privacy and data control as crucial key points to enable data sovereignty in terms of PHRs. This term is defined by analysing the GDPR and how local policies extend it. We showed that data sovereignty requires more than what is demanded by current laws and their execution with the example of information self-determination, which is considered a fundamental right in Germany. From a technical point of view we see that this concept needs to be extended with a rich digital consent management and ways to enforce automatic data usage tracking for transparency. It has to be mentioned that the addition and extensions can not be purely technical and an interdisciplinary approach is required. Our overview looked at the e-Health policies of the G7 nations with a focus on privacy and patient involvement. The overview showed that while nearly every country has one or more laws that enforces privacy and patient involvement, the implementation is often limited. There are also clear examples proving that privacy and patient involvement is a key factor for the acceptance of a PHR. With those results we define a data sovereignty framework for PHRs. We propose that broad access, fine granular control, informed decision making, intuitive user experience and comprehensible transparency can help to give a user control over his data and establishes trust in the PHR system.

Our results show that more work is required in various fields to define a good framework for patient involvement and data sovereignty for PHRs. This should be done in an interdisciplinary effort. On the one hand there is the legal view. Further research and legislation is required to define the guidelines for the underlying technological possibilities of a PHR. In addition the ethical view should also be considered. Questions like what possibilities in terms of data control and privacy should be, if at all, limited for the patient need to be investigated. The issue when dealing with the trade off between usability and pre-filled decisions should be evaluated from an ethical point of view. Finally, besides the general technological implementation, there needs to be user research to create a proper user experience. This should investigate how a patient can be empowered, so he can understand his decision making and data control

without being overwhelmed by too many possibilities. Besides those open questions further observation of the current development in e-Health worldwide is necessary. One thing that could lead to new data on patient acceptance will be the upcoming launch of the German ePA. With a lot of controversy in terms of data protection and security, it remains to be seen how the compromises in those matters affect the adaption and acceptance of the project. While this is just one example of on-going digitisation, there will be more data and studies that could be used to refine our presented framework.

References

[1] S Bologna et al. “Electronic Health Record in Italy and Personal Data Protection.” In: *Eur J Health Law* 23.3 (2016), pp. 265–277.

[2] German Bundesrat. *German e-Health Law*. URL: http://dipbt.bundestag.de/dip21/brd/2015/0257-15.pdf.

[3] Philippe Burnel. “The introduction of electronic medical records in France: More progress during the second attempt”. In: *Health Policy* 122.9 (2018), pp. 937–940.

[4] Government of Canada. *Personal Information Protection and Electronic Documents Act*. URL: https://laws-lois.justice.gc.ca/ENG/ACTS/P-8.6/index.html.

[5] Tobias Dehling and Ali Sunyaev. “Secure provision of patient-centered health information technology services in public networks—leveraging security and privacy features provided by the German nationwide health information technology infrastructure”. In: *Electronic Markets* 24.2 (2014), pp. 89–99.

[6] Department of Health and Social Care. *The power of information: giving people control of the health and care information they need*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/213689/dh_134205.pdf. Accessed 25 November 2020. 2012.

[7] Jos Dumortier and Griet Verhenneman. “Legal Regulation of Electronic Health Records: A Comparative Analysis of Europe and the US”. In: *eHealth: Legal, Ethical and Governance Challenges*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 25–56.

[8] European Commission. *e-Health: strategy and ongoing programs*. https://ec.europa.eu/health/sites/health/files/ehealth/docs/ev_20151123_co06_en.pdf. Accessed 25 November 2020. 2015.

[9] European Commission. *Overview of the national laws on electronic health records in the EU Member States*. https://ec.europa.eu/health/ehealth/projects/nationallaws_electronichealthrecords_de. Accessed 25 November 2020. 2016.

[10] European Commission. *Overview of the national laws on electronic health records in the EU Member States: National Report for Italy*. https://ec.europa.eu/health/sites/health/files/ehealth/docs/laws_italy_en.pdf. Accessed 25 November 2020. 2014.

[11] gematik. *Whitepaper Datenschutz*. https://www.gematik.de/fileadmin/user_upload/gematik/files/Publikationen/gematik_Whitepaper-Datenschutz_web_202009.pdf. Accessed 25 November 2020; In German. 2020.

[12] J Hoeksma. “The NHSs care.data scheme: what are the risks to privacy”. In: *BMJ* 348 (2014), g1547.

[13] Ministère de l’Économie des Finances et de la Relance. *France numérique 2012-2020*. https://www.economie.gouv.fr/files/files/import/2011_france_numerique_consultation/2011_francenumerique2020objectifs.pdf. Accessed 25 November 2020. 2011.

[14] K Morris et al. “Designing an Authorization System Based on Patient Privacy Preferences in Japan.” In: *Stud Health Technol Inform* 247 (2018), pp. 71–75.

[15] Otake, Tomoko. *Medical big data to be pooled for disease research and drug development in Japan*. https://www.japantimes.co.jp/news/2017/05/15/reference/medical-big-data-pooled-disease-research-drug-development-japan/. Accessed 25 November 2020. 2017.

[16] World Health Organization (WHO). *Directory of eHealth policies*. https://www.who.int/goe/policies/countries/en/. Accessed 25 November 2020. 2019.

Characterization of Mueller matrices in retroreflex ellipsometry

Chia-Wei Chen

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
chia-wei.chen@kit.edu

Abstract

Ellipsometry is a widely-used technique for characterizing materials and thin films. The principle is based on the polarization changes after light is reflected or transmitted at a sample. In general, the shape of the sample should be flat or nearly flat because ellipsometry is sensitive to the angle of incidence, tilt angle and the sample position (height variation). For nonplanar surfaces, retroreflex ellipsometry was proposed to solve the problem of the alignment. Despite of the Mueller matrix, the coherency matrix is often used for depolarization and noise reduction. In retroreflex ellipsometry, the measured Mueller matrix can be seen as a dual-rotation transformation. Therefore, it is important to discuss the changes of reference frames for Mueller matrices. In this report, the polarization model of retroreflex ellipsometry will be introduced. Decompositions and invariant quantities of a Mueller matrix with a dual-rotation transformation will be discussed.

1 Introduction

Ellipsometry is a widely-used technique for characterizing materials and thin films, e.g., in the semiconductor industry, biology and nanotechnology. The prin-

ciple is based on the polarization changes after light is reflected or transmitted at a sample. The polarization characteristics can be described by Fresnel equations. The advantages of ellipsometry are non-destructive, fast measurements, and high accuracy and sensitivity. In general, the geometric shape of samples should be flat in order to fulfill the law of reflection or Snell's law. For nonplanar samples, the curvatures of the surface alter the reflected or transmitted light which causes experimental errors due to the misalignment. The worst-case scenario is that the detector cannot receive any signal. This restriction limits the feasibility of in-line measurements for industrial applications. In the last two decades, many approaches were proposed to overcome the shape restriction [19, 15, 10, 26, 14, 23, 16, 8]. However, these studies have some constraints, e.g., small measurement ranges, a short working distance, and complicated system design. In order to conquer these drawbacks, the concept of retroreflex ellipsometry

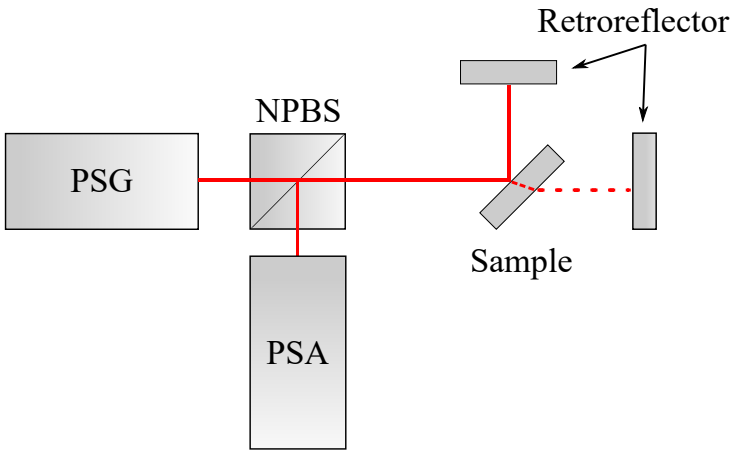


Figure 1.1: Schematic of the retroreflex ellipsometer in the reflection and transmission configurations, showing the polarization state analyzer (PSA), polarization state generator (PSG) and non-polarizing beam-splitter (NPBS).

(RRE) has been proposed at Fraunhofer IOSB[13, 4, 5, 18, 17]. Figure 1.1 shows the configuration of RRE whose concept is based on return-path ellipsometry [20, 2]. The key element in RRE is a retroreflector which returns the light

along the same path and has the same polarization characteristics as an ideal mirror regardless of the angle of incidence (within an angular range up to $\pm 30^{\circ}$). Therefore, the alignment of angles and position between the detector and the sample is automatically achieved. In this paper, the polarization model of retroreflex ellipsometry will be introduced, and decompositions and invariant quantities of the measured Mueller matrices will be discussed.

2 Polarization model of retroreflex ellipsometry

The polarization characteristics of optical elements or the interaction at the boundaries can be described by Jones vectors $\mathbf{E}$, Jones matrices $\mathbf{J}$, Stokes vectors $\mathbf{S}$ and Mueller matrices $\mathbf{M}$ [3, 9]. Jones vectors and Jones matrices can only be used for completely polarized light and nondepolarizing systems while Stokes vectors and Mueller matrices can be used for partially polarized light and depolarizing systems. A Jones matrix can be converted to the Mueller matrix by the transformation:

$$\mathbf{M} = \mathbf{A}(\mathbf{J} \otimes \mathbf{J}^{*})\mathbf{A}^{-1}, \tag{2.1}$$

where $\otimes$ denotes the Kronecker product, the asterisk denotes complex conjugation, and $\mathbf{A}$ is the transformation matrix given by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & i & -i & 0 \end{bmatrix}. \tag{2.2}$$

Stokes vectors $\mathbf{S}$ (4×1 vector) describe the polarization state of the electromagnetic waves including fully polarized, partially polarized, or unpolarized light. Mueller matrices $\mathbf{M}$ (4×4 matrix) characterize the interaction between mediums and polarized light.

$$\mathbf{S} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix}, \mathbf{M} = \begin{bmatrix} m_{00} & m_{01} & m_{02} & m_{03} \\ m_{10} & m_{11} & m_{12} & m_{13} \\ m_{20} & m_{21} & m_{22} & m_{23} \\ m_{30} & m_{31} & m_{32} & m_{33} \end{bmatrix} \tag{2.3}$$

The Mueller matrix of an isotropic sample $\mathbf{M}_S$ can be expressed by the NSC representation:

$$\mathbf{M}_S = \begin{bmatrix} 1 & -N & 0 & 0 \\ -N & 1 & 0 & 0 \\ 0 & 0 & C & S \\ 0 & 0 & -S & C \end{bmatrix}, \tag{2.4}$$

where $N = \cos 2\Psi$, $S = \sin 2\Psi \sin \Delta$, and $C = \sin 2\Psi \cos \Delta$. Ψ and Δ, which are functions of the angle of incidence and the refractive index of the sample, represent amplitude ratio and phase difference. When Mueller matrices are presented with different coordinate frames, the coordinate transformation should be applied. The Mueller matrix of a coordinate rotation $\mathbf{M}_R(\alpha)$ can be described as

$$\mathbf{M}_R(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\alpha & -\sin 2\alpha & 0 \\ 0 & \sin 2\alpha & \cos 2\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.5}$$

where α is the rotation angle.

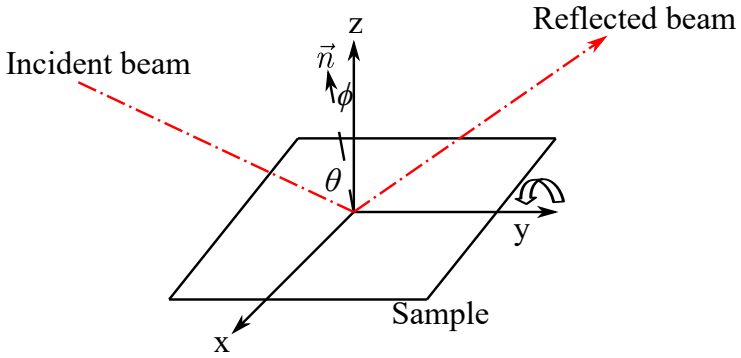


Figure 2.1: Definitions for the angle of incidence θ and the tilt angle ϕ of a tilted sample rotated around the y-axis.

Figure 2.1 shows that a flat sample on the x-y plane rotates around y-axis. If the surface is in parallel to x-y plane, the surface normal is the z-axis and the

plane of incidence is determined by the z-axis and the incident beam. When the sample rotates around the y-axis, the surface normal of the sample becomes the vector $\vec{n}$. The tilt angle ϕ is defined by the surface normal $\vec{n}$ and the z-axis, and the angle of incidence θ is determined by the surface normal $\vec{n}$ and the incident beam. This model can be extended to nonplanar surfaces. A nonplanar surface can be seen as a flat surface rotates around y-axis. The incident angle θ and the tilt angle ϕ define the surface normal. The Mueller matrix model of the tilt sample for RRE is shown as [14]

$$\mathbf{M}_R(\alpha_{\mathrm{PSA}}) \cdot \mathbf{M}_S \cdot \mathbf{M}_{Retroreflector} \cdot \mathbf{M}_S \cdot \mathbf{M}_R(\alpha_{\mathrm{PSG}}), \tag{2.6}$$

where α_{PSA} and α_{PSG} are the rotation angles of the polarization state analyzer (PSA) and the polarization state generator (PSG). Assuming that the system is perfectly aligned, we can use the relation $\mathbf{M}_R(\alpha_{\mathrm{PSA}}) = \mathbf{M}_R(\alpha_{\mathrm{PSG}}) = \mathbf{M}_R(\phi)$ to simplify the equation as

$$\mathbf{M}_2 = \mathbf{M}_R(\phi) \cdot \mathbf{M}_1 \cdot \mathbf{M}_R(\phi), \tag{2.7}$$

where $\mathbf{M}_1 = \mathbf{M}_S \cdot \mathbf{M}_{Retroreflector} \cdot \mathbf{M}_S$. The ellipsometric parameters (Ψ, Δ) and the tilt angle ϕ from the Mueller matrix $\mathbf{M}_2$ can be solved by a numerical fitting method. $\mathbf{M}_2$ can be seen as a dual-rotation transformation of $\mathbf{M}_1$[11].

3 Decompositions of Mueller matrices with dual-rotation transformations

Figure 3.1 shows the different domains of 4×4 matrices. Mueller matrices are a subset of real 4×4 matrices because Mueller matrices contain physical properties (polarization). Mueller-Jones matrices are matrices which are derivable from Jones matrices. Therefore, not all Mueller matrices can be transformed from Jones matrices because Jones matrices only deal with nondepolarized systems. There have been many studies discussing necessary and sufficient conditions for a Mueller matrix [6, 21]. A Mueller matrix is a linear transformation of Stokes vectors. Hence, Mueller matrices must fulfill Stokes criterion:

$$s_0 \geq (s_1^2 + s_2^2 + s_3^2)^{\frac{1}{2}}. \tag{3.1}$$

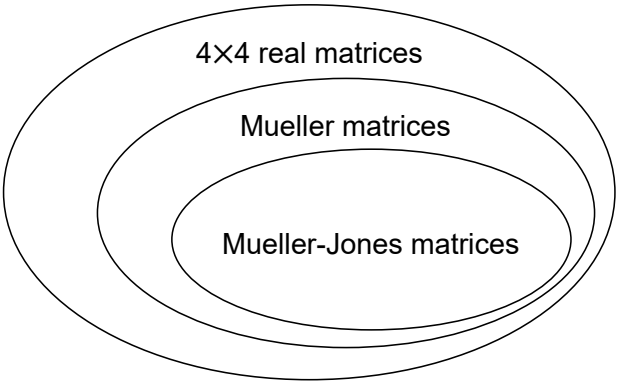


Figure 3.1: Different domains for 4×4 matrices.

For every Stokes vector $\mathbf{S}$ satisfies the criterion, the product of the Mueller matrix $\mathbf{M}$ and Stokes vector $\mathbf{S}$ also satisfies the criterion. Then the Mueller matrix fulfills the Stokes criterion.

In order to analyze depolarization, the wave coherency matrix $\mathbf{\Phi}$ is proposed as [24]:

$$\mathbf{\Phi} = \mathbf{E}(t) \otimes \mathbf{E}(t)^* = \begin{bmatrix} E_x(t)E_x^*(t) & E_x(t)E_y^*(t) \\ E_y(t)E_x^*(t) & E_y(t)E_y^*(t) \end{bmatrix} = \begin{bmatrix} j_{00} & j_{01} \\ j_{10} & j_{11} \end{bmatrix}, \tag{3.2}$$

where $\mathbf{E(t)}$ is a quasi-monochromatic wave whose amplitudes and phases depend on the time t. We can see depolarisation is related to second order products of the quasi-monochromatic wave. This concept can be applied to Mueller matrices. The covariance matrix $\mathbf{H}$ is defined as Kronecker product of the corresponding Jones covariance vector [22]:

$$\mathbf{H} = \frac{1}{2}\mathbf{T} \otimes \mathbf{T}^*, \tag{3.3}$$

where $\mathbf{T} = [j_{00} \;\; j_{01} \;\; j_{10} \;\; j_{11}]^T$ and j_{ij} is the element of the 2×2 Jones matrix. Hence, $\mathbf{H}$ is a 4×4 matrix. It is obvious that $\mathbf{H}$ and $\mathbf{M}$ are linearly related. Therefore, $\mathbf{H}$ can be written in terms of the elements m_{ij} of $\mathbf{M}$ as:

$$\mathbf{H}(\mathbf{M}) = \frac{1}{2}\begin{bmatrix} m_{00}+m_{01} & m_{02}+m_{12} & m_{20}+m_{21} & m_{22}+m_{33} \\ +m_{10}+m_{11} & +i(m_{03}+m_{13}) & -i(m_{30}+m_{31}) & +i(m_{23}-m_{32}) \\ m_{02}+m_{12} & m_{00}-m_{01} & m_{22}-m_{33} & m_{20}-m_{21} \\ -i(m_{03}+m_{13}) & +m_{10}-m_{11} & -i(m_{23}+m_{32}) & -i(m_{30}-m_{31}) \\ m_{20}+m_{21} & m_{22}-m_{33} & m_{00}+m_{01} & m_{02}-m_{12} \\ +i(m_{30}+m_{31}) & +i(m_{23}+m_{32}) & -m_{10}-m_{11} & +i(m_{03}-m_{13}) \\ m_{22}+m_{33} & m_{20}-m_{21} & m_{02}-m_{12} & m_{00}+m_{01} \\ -i(m_{23}-m_{32}) & +i(m_{30}-m_{31}) & -i(m_{03}-m_{13}) & +m_{10}+m_{11} \end{bmatrix} \tag{3.4}$$

The covariance matrix $\mathbf{H}$ provides necessary and sufficient conditions for a Mueller matrix to be derivable from a Jones matrix [1]. The form of $\mathbf{H}$ is a positive semidefinite Hermitian matrix, which means its eigenvalues are non-negative. In other words, a matrix is a physical realizable Mueller matrix if its coherency matrix $\mathbf{H}$ has non-negative eigenvalues. This concept can be used to determine physical Mueller matrices and reduce experimental errors [7].

Experimental errors in ellipsometry might induce nonphysical Mueller matrices (negative eigenvalues in corresponding covariance matrices $\mathbf{H}$). For example, a depolarizing Mueller matrix is measured due to the noise from the light source and the detector. The idea of sum decomposition or matrix filtering for experimental Mueller matrices is proposed by Cloude [7]. The covariance matrix of a physically realizable Mueller matrix can be decomposed to four covariance matrices of Mueller-Jones matrices as:

$$\mathbf{H} = \lambda_1\mathbf{H}_1 + \lambda_2\mathbf{H}_2 + \lambda_3\mathbf{H}_3 + \lambda_4\mathbf{H}_4. \tag{3.5}$$

If a Mueller matrix is nonphysical, at least one eigenvalue of its covariance matrix is negative. The filtering concept is to remove any negative contributions and covert the remaining term to a Mueller matrix, which can be described as:

$$\sum_{i=1}^{4}\frac{1}{2}(1+\mathrm{sgn}(\lambda_i))\mathbf{H}_i \Rightarrow \mathbf{M}_{filtering}, \tag{3.6}$$

where sgn is the sign function and λ_i is the eigenvalue of $\mathbf{H}$. Finally, The nearest non-depolarising Mueller matrix is obtained. This method is proved as an optimal filtering [25].

We can apply covariance matrices in retroreflex ellipsometry. The Mueller matrix for a gold sample at a wavelength of 632.8 nm and an incident angle of

70° is given by

$$\mathbf{M}_{gold} = \begin{bmatrix} 1 & -0.094 & 0 & 0 \\ -0.094 & 1 & 0 & 0 \\ 0 & 0 & 0.802 & -0.589 \\ 0 & 0 & 0.589 & 0.802 \end{bmatrix}. \tag{3.7}$$

The eigenvalues of the covariance matrix $\mathbf{H}(\mathbf{M_{gold}})$ are $[1, 0, 0, 0]$, which means the matrix $\mathbf{M}_{gold}$ is a Mueller-Jones matrix. When the gold sample is tilted with 5°. The Mueller matrix $\mathbf{M}_{gold}$ becomes

$$\mathbf{M}'_{gold} = \begin{bmatrix} 1 & -0.093 & -0.016 & 0 \\ -0.093 & 0.946 & 0.309 & -0.102 \\ 0.016 & -0.308 & 0.748 & -0.580 \\ 0 & -0.102 & 0.580 & 0.802 \end{bmatrix} \tag{3.8}$$

We can observe that the off-diagonal 2×2 blocks are nonzero elements. The eigenvalues of the covariance matrix $\mathbf{H}(\mathbf{M'_{gold}})$ are $[1, 0.015, -0.015, 0]$. Minus eigenvalue means that $\mathbf{M}'_{gold}$ is not a physically realizable Mueller matrix. The change of plane of incidence caused the anisotropic and depolarizing effect.

We can prove the Mueller matrix with a dual-rotation transformation $\mathbf{M}_2$ is not positive semi-definite by Sylvester's criterion [12]. A Hermitian matrix is positive semi-definite if and only if all principal minors of it are non-negative. For a 4×4 matrix, there are 15 principal minors D_k, where k is the order. The principal minors of $\mathbf{M}_2$ is shown as

$$\begin{aligned} D_1 &= [0, \frac{1}{2}(1-C)(1-\cos 4\phi), \frac{1}{2}(1-C)(1+\cos 4\phi), 1+C] \\ D_2 &= [0, 0, 0, 0, 0, 0] \\ D_3 &= [0, 0, 0, \frac{1}{2}S^2(1-C)(-1+\cos 8\phi)] \\ D_4 &= 0 \end{aligned} \tag{3.9}$$

Since $C, S \in [-1, 1]$ and $\phi \in [-90^\circ, 90^\circ]$, the principal minor in D_3 is negative when the tilt angle is not zero. The special cases are $S = 0$ and $C = 1$. $S = 0$ means $\Psi = 45^\circ$ and the corresponding Mueller matrix is an ideal depolarizer which output randomly polarized light. $C = 1$ means $\Psi = 45^\circ$ and $\Delta = 0^\circ$ or

360° and the Mueller matrix of this case is the same as the Mueller matrix of air. Except of these two cases, we can use this property to find the tilt angle by maximizing the $h_{fidelity}$ index[7]:

$$h_{fidelity} = 10\log\frac{\sum|\lambda_-|}{\sum\lambda_+}, \tag{3.10}$$

where λ_+ and λ_- are positive and negative eigenvalues of the correspingding covariance matrix.

4 Invariant quantities of a Mueller matrix with a dual-rotation transformation

There are polarimetric quantities which keep invariant under reference frame rotations. These invariant quantities can be used for determination of orientations of anisotropic materials and provide physical information. The Muller matrix model of a sample with a tilt angle in retrorelfex ellipsometry can be seen as a Mueller matrix with a dual-rotation transformation. The full form of $\mathbf{M}_1$ and $\mathbf{M}_2$ are expressed as:

$$\mathbf{M}_1 = \begin{bmatrix} 1+N^2 & -2N & 0 & 0 \\ -2N & 1+N^2 & 0 & 0 \\ 0 & 0 & S^2-C^2 & -2CS \\ 0 & 0 & 2CS & S^2-C^2 \end{bmatrix}, \tag{4.1}$$

$$\mathbf{M}_2 = \begin{bmatrix} 1+N^2 & -2N\cos 2\phi & 2N\sin 2\phi & 0 \\ -2N\cos 2\phi & 1-S^2+(1-C^2)\cos 4\phi & -(1-C^2)\sin 4\phi & 2CS\sin 2\phi \\ -2N\sin 2\phi & (1-C^2)\sin 4\phi & S^2-1+(1-C^2)\cos 4\phi & -2CS\cos 2\phi \\ 0 & 2CS\sin 2\phi & 2CS\cos 2\phi & S^2-C^2 \end{bmatrix}. \tag{4.2}$$

Compared $\mathbf{M}_1$ with $\mathbf{M}_2$, the following parameters are rotation invariant:

$$m_{00}, m_{03}, m_{30}, m_{33}, \tag{4.3}$$

$$m_{00}^2+m_{02}^2, m_{10}^2+m_{20}^2, m_{13}^2+m_{23}^2, m_{31}^2+m_{32}^2, \tag{4.4}$$

$$m_{11}^2+m_{12}^2+m_{21}^2+m_{22}^2, \tag{4.5}$$

$$\mathrm{Det}(\mathbf{M}), \tag{4.6}$$

$$\lambda_1 + \lambda_2 + \lambda_3 - 3\lambda_4, \tag{4.7}$$

where Det denotes determinant and λ_i is the eigenvalue of the correspingding covariance matrix.

From $\mathbf{M}_2$, the tilt angle ϕ can be obtained by

$$\phi = \tan^{-1}\frac{m_{02}}{m_{01}} = \tan^{-1}\frac{m_{20}}{m_{10}} = \tan^{-1}\frac{m_{13}}{m_{23}} = \tan^{-1}\frac{m_{31}}{m_{32}}. \tag{4.8}$$

The NSC parameters can be determined by

$$\begin{aligned} N^2 &= \frac{m_{01}^2 + m_{02}^2}{4} = \frac{m_{10}^2 + m_{20}^2}{4} \\ S^2 &= \frac{1}{2}(\sqrt{m_{31}^2 + m_{32}^2 + m_{33}^2} + m_{33}) = \frac{1}{2}(\sqrt{m_{13}^2 + m_{23}^2 + m_{33}^2} + m_{33}) \\ C^2 &= \frac{1}{2}(\sqrt{m_{31}^2 + m_{32}^2 + m_{33}^2} - m_{33}) = \frac{1}{2}(\sqrt{m_{13}^2 + m_{23}^2 + m_{33}^2} - m_{33}). \end{aligned} \tag{4.9}$$

Finally, the ellipsometric parameters Ψ and Δ can be determined by the NSC parameters. It is worthwhile to mention that the NSC parameters can be caulculated without knowing the tilt angle ϕ. In other words, NSC parameters are only related to the angle of incidence and material properties. If the refractive index of the sample is known, the angle of incidence can be solved analytically.

For isotropic materials, tilt angles induce anisotropic Mueller matrices. There are variant and invariant polarimetric quantities in the anisotropic matrices. These invariant quantites provide the information of rotation of reference frames. Moreover, the invariant quantities can be extended to anisotropic materials for separating azimuthal orientation and tilt angles.

5 Summary

In this report, the principle of retrorefelx ellipsometry, coherency matrix, Cloude's decomposition and invariant quantities for nonplanar surfaces have been introduced. The concept of RRE can measure samples with nonplanar

shapes. The retroreflector acts as an ideal mirror regardless the incident angle ($< 30°$). The polarization model of the tilt sample can be seen as a dual-rotation transformation and the form of an isotropic Mueller matrix after rotation becomes anisotropic. The coherency matrix $\mathbf{H}$ provides necessary and sufficient conditions for a Mueller matrix which can be derivable from a Jones matrix. The Cloude's decomposition can reduce the experimental noise by filtering nonphysical contribution (negative eigenvalues) and can also be used to determine the tilt angle ϕ. While the sample has a tilt angle, $\mathbf{H}$ becomes nonphysical. This nonphysical Mueller matrix provides indication of tilt angles. Compared to the physical matrix (without tilting) and nonphysical matrix (with tilting), invariant quantities provide another way to calculate tilt angles and ellipsometric parameters. In the future, we plan to use these properties of Mueller matrices to improve the procedure of determination of the tilt angle ϕ and ellipsometric parameters (Ψ, Δ) and extend this method to anisotropic materials.

References

[1] Donald GM Anderson and Richard Barakat. "Necessary and sufficient conditions for a Mueller matrix to be derivable from a Jones matrix". In: *JOSA A* 11.8 (1994), pp. 2305–2319.

[2] R.M.A. Azzam. "Return-path Ellipsometry and a Novel Normal-incidence Null Ellipsometer (NINE)". In: *Optica Acta: International Journal of Optics* 24.10 (1977), pp. 1039–1049. ISSN: 0030-3909. DOI: 10.1080/713819411.

[3] Rasheed Mohammed Abdel-Gawad Azzam and Nicholas Mitchell Bashara. *Ellipsometry and polarized light*. 4. impression, paperback ed. North-Holland personal library. Amsterdam: Elsevier, 1999. ISBN: 0-444-87016-4.

[4] Chia-Wei Chen et al. "Measurement of ellipsometric data and surface orientations by modulated circular polarized light / Messung von ellipsometrischen Daten und Oberflächenorientierungen durch moduliertes

zirkular polarisiertes Licht". In: *tm - Technisches Messen* 86.s1 (2019), pp. 32–36. ISSN: 2196-7113. DOI: 10.1515/teme-2019-0047.

[5] Chia-Wei Chen et al. "Retroreflex ellipsometry for isotropic substrates with nonplanar surfaces". In: *Journal of Vacuum Science & Technology B* 38.1 (2020), p. 014005. DOI: 10.1116/1.5121854. eprint: https://doi.org/10.1116/1.5121854. URL: https://doi.org/10.1116/1.5121854.

[6] Shane Cloude. *Polarisation: applications in remote sensing*. OUP Oxford, 2009.

[7] Shane R Cloude. "Conditions for the physical realisability of matrix operators in polarimetry". In: *Polarization Considerations for Optical Systems II*. Vol. 1166. International Society for Optics and Photonics. 1990, pp. 177–187.

[8] Matthias Duwe et al. "Thin-film metrology of tilted and curved surfaces by imaging Mueller-matrix ellipsometry". In: *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 37.6 (2019), p. 062908. ISSN: 2166-2746. DOI: 10.1116/1.5122757.

[9] Hiroyuki Fujiwara. *Spectroscopic ellipsometry: Principles and applications*. Chichester, England and Hoboken, NJ: John Wiley & Sons, 2007. ISBN: 9780470060186. DOI: 10.1002/9780470060193.

[10] Abhijeet Ghosh et al. In: *ACM T. Graphic* 29.6 (2010), p. 162. ISSN: 4503-0439. DOI: 10.1145/1866158.1866163.

[11] José J Gil. "Invariant quantities of a Mueller matrix under rotation and retarder transformations". In: *JOSA A* 33.1 (2016), pp. 52–58.

[12] George T Gilbert. "Positive definite matrices and Sylvester's criterion". In: *The American Mathematical Monthly* 98.1 (1991), pp. 44–46.

[13] Matthias Hartrumpf et al. In: *Tech. Mess.* (2019). [published online ahead of print Sep. 6, 2019]. ISSN: 2196-7113. DOI: 10.1515/teme-2019-0097.

[14] Blaine Johs and Ping He. “Substrate wobble compensation for in situ spectroscopic ellipsometry measurements”. In: *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 29.3 (2011), p. 03C111. ISSN: 2166-2746. DOI: 10.1116/1.3555332.

[15] Kan Yan Lee and Yu Faye Chao. “The Ellipsometric Measurements of a Curved Surface”. In: *Japanese Journal of Applied Physics* 44.7L (2005), p. L1015. ISSN: 1347-4065. DOI: 10.1143/JJAP.44.L1015. URL: http://iopscience.iop.org/article/10.1143/JJAP.44.L1015/pdf.

[16] Weiqi Li et al. “Characterization of curved surface layer by Mueller matrix ellipsometry”. In: *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 34.2 (2016), p. 020602. ISSN: 2166-2746. DOI: 10.1116/1.4943952.

[17] Christian Negara, Thomas Längle, and Jürgen Beyerer. “Analytic solutions for calculating the surface inclination of isotropic media and bare substrates by using reflection-based generalized ellipsometry”. In: *Journal of Vacuum Science & Technology B* 38.3 (2020), p. 034012. DOI: 10.1116/1.5144506. URL: https://doi.org/10.1116/1.5144506.

[18] Christian Negara, Thomas Längle, and Jürgen Beyerer. “Imaging ellipsometry for curved surfaces”. In: *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 38.1 (2020), p. 014016.

[19] Ulrich Neuschaefer-Rube and Wolfgang Holzapfel. “Simultaneous measurement of surface geometry and material distribution by focusing ellipsotopometry”. In: *Applied Optics* 41.22 (2002), p. 4526. ISSN: 0003-6935. DOI: 10.1364/AO.41.004526.

[20] H. M. O’Bryan. “The Optical Constants of Several Metals in Vacuum*”. In: *JOSA* 26.3 (1936), pp. 122–127. DOI: 10.1364/JOSA.26.000122. URL: https://www.osapublishing.org/viewmedia.cfm?uri=josa-26-3-122&seq=0.

[21] Jose Jorge Gil Perez and Razvigor Ossikovski. *Polarized light and the Mueller matrix approach*. CRC press, 2016.

[22] R Simon. “The connection between Mueller and Jones matrices of polarization optics”. In: *Optics Communications* 42.5 (1982), pp. 293–297.

[23] Toshihide Tsuru. In: *Opt. Express* 21.5 (2013), pp. 6625–6632. DOI: 10.1364/OE.21.006625.

[24] Emil Wolf. “Coherence properties of partially polarized electromagnetic radiation”. In: *Il Nuovo Cimento (1955-1965)* 13.6 (1959), pp. 1165–1181.

[25] Tim Zander and Juergen Beyerer. “Mueller matrix cone and its application to filtering”. In: *OSA Continuum* 3.6 (2020), pp. 1376–1384.

[26] Yi Zhang. In: *Rev. Sci. Instrum.* 81.8 (2010), p. 085101. DOI: 10.1063/1.3465314.

A Data Annotation Process for Human Activity Recognition in Public Places

Mickael Cormier

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

Abstract

Behavior analysis of individuals in crowds or groups of people in public places through surveillance cameras gains importance for several different actors. Automatically detecting and understanding pedestrians in real-world uncooperative scenarios is very challenging. Common issues such as limited annotated data, unreliable data and annotation quality, and appropriate use of this data for supervised learning often originate in steps preceding the modeling of specialized neural network architectures. In this report, the necessity and requirements for designing a reliable data annotation process are presented. Some precise ideas for automation through neural networks are discussed in a conceptual manner.

1 Introduction

Automated analysis of video content through deep learning algorithms offers tremendous potential for autonomous driving, health care, agriculture, surveillance for both home and public places. In the last years most annotated datasets required by such algorithms have been labeled manually. ImageNet [6], the

first major publicly available large scale dataset was published in 2009 with more than 14 millions hand-annotated images, for which 20,000 to 30,000 people a year worked for several years through crowd-sourcing [14]. Systems which could potentially threaten human lives, such as assisting systems for surgery or autonomous driving require reliable, high quality data, with strict and explainable validation measures. Furthermore the data need to cover rare events and scenarios which could represent critical safety issues. For instance, to ensure the correct functioning of a sensor system for autonomous driving, it must be demonstrated over a distance of 300,000 km within a given scenario, meaning over 240 millions frames and 3.6 billions objects are to be annotated accurately [20]. Considering an average of 60 seconds for the thorough annotation of an object, the sole creation of a single validation dataset for a given scenario represents a several-year project with hundreds of full-time annotators.

Similarly, safety related surveillance projects not only require detection of people but also an estimation of their body posture as well as their activity. Given a public place covered by several surveillance cameras, scenes of interest need to be annotated frame by frame, with an average of 50 persons in sequences recorded with 30 frames per seconds for several minutes. Considering several camera clusters, manual annotation appears almost not affordable regarding the rigorous and transparent validation of the system which are required to fulfill legal and ethical requirements. Therefore, the average annotation time per object must be reduced. To this aim a highly automated data annotation process and data system is enquired.

The remainder of this report is oriented toward the design of such a process and, thus, organized as follows: the concept of data annotation and automation are presented in Section 2, requirements of a reliable annotation process are introduced in Section 3, while the steps of the data annotation process and the data quality assurance are discussed in Section 3.2 and Section 4 respectively. A conclusion is given in Section 5.

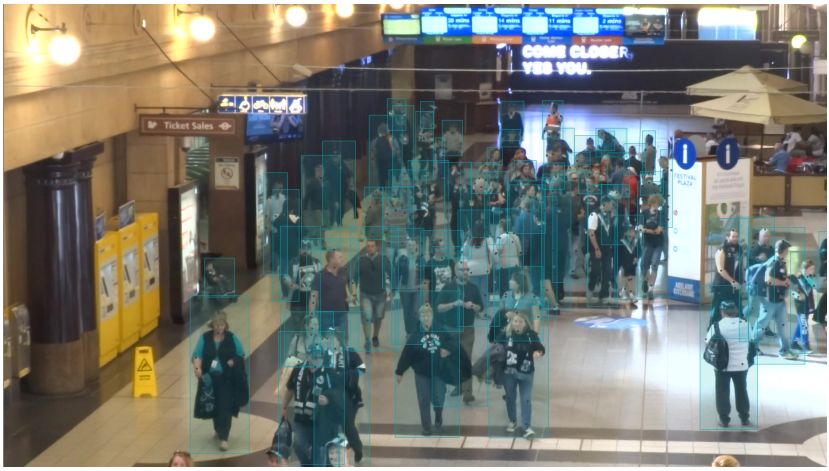


Figure 2.1: Manual keypoint and bounding box annotations for a crowded scene in [5].

2 Data Annotation and Automation

In this work we consider the annotation process for human activity recognition in public places such as a public transportation hub. Cameras are placed strategically in order to monitor traffic, assure the traveler's safety and security. Therefore, large areas coverage using clusters of cameras with wide field of views offering multiple views of given hotspots is required. Thus, data for raw scenes including multiple views and dozens up to larger crowds of hundreds of pedestrians result from this setup, as illustrated in 2.1.

An important part of use cases in the field of human activity recognition is based on supervised learning, which means training data with target annotation is required in order to update a prediction model. The process of collecting annotation is often much more expensive than the process of collecting the data itself. This annotated data, the ground truth, is however limited by the expertise of the person annotating. Depending on the complexity of the task and its specifications, a label may require on the one hand pixel-wise precision, e.g. semantic segmentation, keypoint detection, precise 2d and 3d bounding

boxes. On the other hand, tasks such as person tracking, re-identification and temporal action localization in video require a more profound understanding of the scene developing. Furthermore, several different annotation formats may internally define the same type of annotation differently [13, 2, 8, 10, 22, 4, 17] resulting in a supplementary layer of complexity for the annotator, if not intuitive or usual. For instance a 2d bounding box in COCO [13] is defined as a (x-top left, y-top left, width, height) tuple, while a Pascal VOC [8] bounding box is defined by the tuple (x-top left, y-top left,x-bottom right, y-bottom right) and YOLO [19] defines a bounding box by it's center. Annotation software such as [22, 8, 3] support multiple formats and offer different tools in order to partly provide automation for specific tasks. However, those are not specifically designed for multiple-views use cases and provide automation only to a limited level.

In the following, we formally define different levels of automation for multiple-view data annotation for human activity recognition in public places.

- **Level 0: No Automation**
  The whole annotation process is done manually. The annotation tools provide simple functions to produce annotation.

- **Level 1: Tool Assistance**
  The annotator is assisted by different tools which minimizes the effort of annotating video frames. Provided multiple views of a the same place, the annotator may switch between views of a sequence while annotating. The annotations are converted to the views of this sequence. Furthermore, given a partly annotated sequence, the tool is able to interpolate the movement of the objects into subsequent frames.

- **Level 2: Partly Automated Annotation**
  Using a point, scribbles or a polygon the annotator defines a region of interest within a frame, the desired annotations are returned by the tool and manually corrected if necessary.

- **Level 3: Highly Automated Annotation**
  The whole (multi-view) sequence is automatically annotated for the chosen

type of labels. The human annotator mainly conducts quality checks and corrections.

- **Level 4: Highly Automated Annotation Pre-Checking**
  The whole (multi-view) sequence is automatically annotated for the chosen type of labels. Quality checks are automatically generated for a human validator to review and extend.

Data annotation is a complex process which comes at a great cost. Designing specific tools which enable automation is an important step in order to produce data annotation at scale. Nevertheless, the quality of the data produced requires constant control and validation. Therefore, a transparent and reliable annotation process is necessary.

3 Designing a Reliable Annotation Process

A reliable annotation process requires clear steps for which different actors, e.g. person or entity, have well defined responsibilities and are held accountable. In this section, a reliable annotation process for human activity recognition in public places is described. The overall process is illustrated in a activity-based flowchart diagram in Figure 3.1. A recent white-paper [7] published during the writing of this work presents similar views regarding an abstract annotation process for supervised learning. In contrast, this report focuses on the concrete case of multi-view annotation processes of crowds and discusses concepts regarding data privacy and ethics. Furthermore, concrete strategies for the use of automatic annotation proposals are defined.

3.1 Participants in the Annotation Process

We identify four entities as participants in the annotation process and define those as follow.

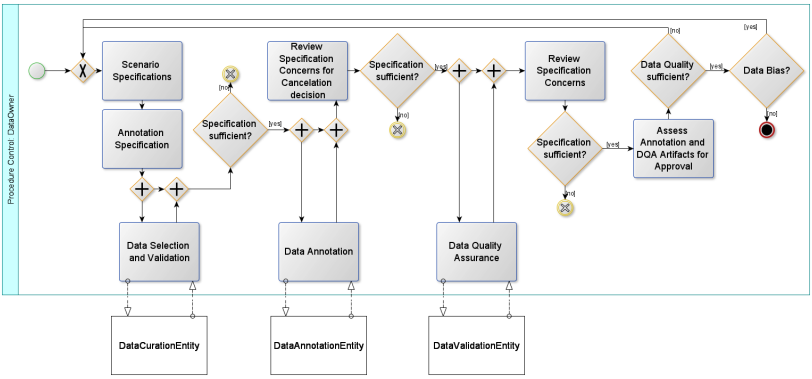


Figure 3.1: Annotation Process Diagram

3.1.1 Data Owner Entity

Ahead of the annotation process the data owner entity is in charge of the data acquisition and identification for the use case, as well as cleaning and preprocessing. The data owner is responsible for the data and the complete annotation process at any time of the process. This entity is in charge of initiating sequence annotation and validation, and finally approve or cancel the result of the process. They are responsible for formulating scenario specifications, quality requirements and identifying scenarios risks and requirements such as anonymization, potential bias, legal issues or underrepresented aspects.

3.1.2 Data Curation Entity

The data curation entity is responsible for assuring compatibility of the data with laws, ethic, potential bias and anonymization requirements. This entity is in charge of analyzing the provided data and scenario specifications, formulate necessary data curation and transformation in order to meet specifications. They extend the scenario specifications with concrete fail condition regarding annotations, such as label balancing or systematic bias.

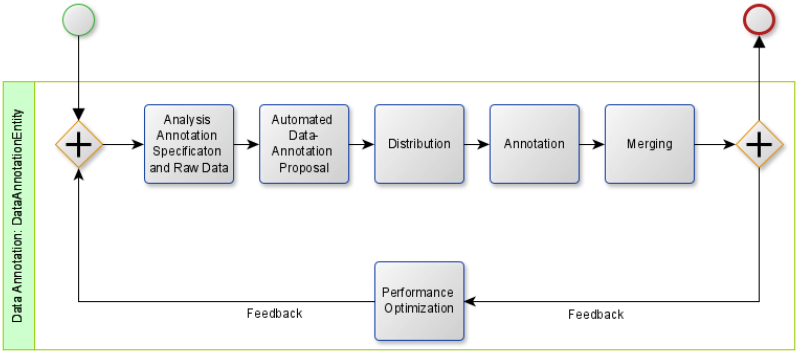


Figure 3.2: Data Annotation Process Diagram

3.1.3 Data Annotation Entity

The data annotation entity produces the annotation and report potential issues either to the curating entity or the data owner. This entity delivers the annotations to the data owner.

3.1.4 Data Validation Entity

The data validation entity fulfills quality assessment of the annotations and reports its results to the data owner or if necessary to the data curation entity.

3.2 Steps of the Annotation Process

In the following the different steps of the annotation process starting from scenario specification to data splitting for training and validating are described. The preceding steps of physical data acquisition and preprocessing are not further discussed here, since those are not directly part of the annotation process itself.

3.2.1 Scenario Specifications

The data owner entity is responsible for the dataset and the annotation project and thus is required to analyse the use case to be covered, its environmental and technical conditions as well as their implication for the machine learning specifications.

Firstly, the conditions of the environment such as lighting, surrounding area variety, weather, time of day of the target scenario have to be defined and openly compared to the provided raw data. The available data should represent the target use case. In case of potential parts missing from the use case, decisions about the addition of supplementary real or synthetic data are required.

Secondly, the technical conditions such as media format, sampling rate, resolution, camera position, potential camera cluster focusing on different places and their implication are analyzed. The aim is to comprehend the feasibility of particular annotation types on specific views, e.g. not enough pixels of the person of interest are provided in order to estimate their pose, too much natural occlusion through fixed construction or vegetation. In case of a camera cluster, it is essential to determine whether the cameras are calibrated and their parameters are known in order to reconstruct 3d representations or convert annotations from one view to another.

Thirdly, while working with data focusing on person in public places important issues emerge concerning privacy, currently applicable laws as well as ethical considerations. Whereas deployed models probably will work on a raw data feed, training and validation data should be anonymized and bio-metrical information of person shouldn't be used implicitly, if not targeting a specific problem related to bio-metrics. Furthermore, the anonymized person should be considered as an avatar of a real person and shouldn't provide unattended long-time information on real-person, e.g. a person may be re-identified in a different view of the same scene whithin the same camera cluster, however this person shouldn't be re-identified in another cluster or in later recordings. Besides data privacy specifications, the data owner entity is expected to address potential bias and ethical issues and proceed carefully during data selection. It is proven that non representative data produce biased results which may impair the quality of the model or, worse, accentuate social inequalities and present bias against ones

ethnicity or gender [1, 15, 18]. Those issues are not only dangerous and noxious, but also damaging to the trustworthiness of the model [11].

Finally the annotation task to be fulfilled is described along it's aim. The intended effect is to increase the implication and engagement of the labeling entity to the task ahead. For instance, the annotation of keypoints for a person may imply that the model to be trained will perform human pose estimation. In this case, the temporal margin of error regarding specific keypoints offers a greater tolerance. However considering pose based action recognition, the lack of precision over time may cause spatial noise over time, which in turn impacts the results of the model.

3.2.2 Annotation Specifications

After analysis of the raw data and the scenario specifications, the data owner entity is expected to define clear and unambiguous labeling specifications regarding which kind of annotation is to be produced, e.g. person detection, instance segmentation, pose estimation. Quality tolerances are precisely defined, e.g. size of bounding box, margin of error tolerated. In dialog with the labeling entity, concrete annotation instructions are derived from these general specifications and must prevent potential issues. Those instructions aim to clearly define solutions for edge cases, clarify unspecific labels and identify potential conflicting instructions. Furthermore, they represent the basis for the annotation validation process and may support the initial configuration of annotation proposal tools. Finally, the specifications and instructions support the potential adaptation of annotation tools if required.

At any time in this step specification concerns may be raised which need to be immediately reviewed, therefore halting the whole process. If the specifications reveal weakness, the process should be aborted and the specifications fixed before starting a new iteration. The earlier such issues are identified, the more efficient the whole process becomes. Despite potential concerns for slowing the process down in its early stages, openly assessing annotation specification is cost efficient and prevents repeated annotation of the same sequences.

3.2.3 Data Selection and Verification for Annotation

The selection of representative data for annotation is carried out according to the annotation specifications. On one hand, depending on the target scenario, the available data is usually subject to subsampling: Spatial sampling, i.e. limit the number of sample per spatial region or completely reject specific spatial regions due to annotation concerns, and/or temporal subsampling, i.e. extract every n-th frame.

On the other hand, a common scenario in the active learning literature represents this data selection step perfectly, which is the unlabeled pool scenario. Considering the yet to be annotated raw data, a round base game is defined. In every round and active learning model ranks the data in the unlabeled pool, and the k bests are selected for annotation, given a fix annotation budget. The selected data is annotated and added to the training set of the active learning model, which is re-trained on the new dataset. These rounds are repeated until the annotation budget is fully drained. Recent methods for CNN not only select the best data for annotation based on efficiency and diversity, but also consider that model training mainly uses batches of data [21, 12, 23].

Both aspects could also be sequentially combined. However, even careful data selection may lead involuntarily to strong biases. This is where the data curation entity is required to formally analyze the selected data and may raise empiric concerns, which need to be reviewed and validated at the end of each annotation process.

3.2.4 Coordination and Distribution

Depending on the volume of selected data, annotation specification, personal availability and prior experiences, the data annotation entity decides on the data and/or label slicing and form of distribution. Overlapping subsets should always be considered for validating individual accuracy as well as for training purposes for new annotators. Different slicing options depending on the data are available. For instance, given multiple labels, it is possible to distribute the target labels between available annotators, e.g. pedestrians, babys, objects and cars for instance segmentation. For motion tracking, sequences could be distributed

per camera, cluster of cameras for scenes. Overall for a dataset composed of multi-view sequences, these may be split in subsequences along the temporal axis and / or the spatial axis. Single subsequences might be shared between the annotator for validation purposes and / or for particular crowded sequences, or different labeling tasks. Eventually, those subsequences are required to be merged in order to finalize the annotation process. In this case merging heuristics are required, e.g. voting, average of all annotation, cherry picking. The merging step may raise issues concerning the annotation specifications, and therefore requiring a new annotation iteration to correct these issues.

3.2.5 Annotation

The data annotation entity performs the labeling and delivers the resulting annotations as illustrated in Figure 3.2. They raise issues immediately during the process and provide feedback on the tool at their disposal for the task.

3.3 Use of Pre-Annotation

Considering the annotation of human poses for crowds in public places with multiple cameras, the annotation effort required increases exponentially. For this reason, the annotation entity relay strongly on automated tools to improve and accelerate the annotation process. As shown in Figure 3.3, efficient tools may improve the annotation greatly.

3.3.1 Use of existing Annotations

Sometimes part of the dataset of an annotation project had been annotated earlier. In this case automated import tools are required to import and reuse these preexisting annotations. However, these annotation are considered (human-) generated pre-annotations pending for review, since they probably do not fully comply the annotation specifications.

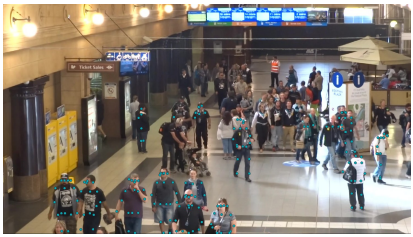
(a) Input Bild (b) Pre-Annotations

Figure 3.3: Pre-Annotations for the domain application greatly improve the quality and speed of the annotations. In this case most of the individuals are already annotated.

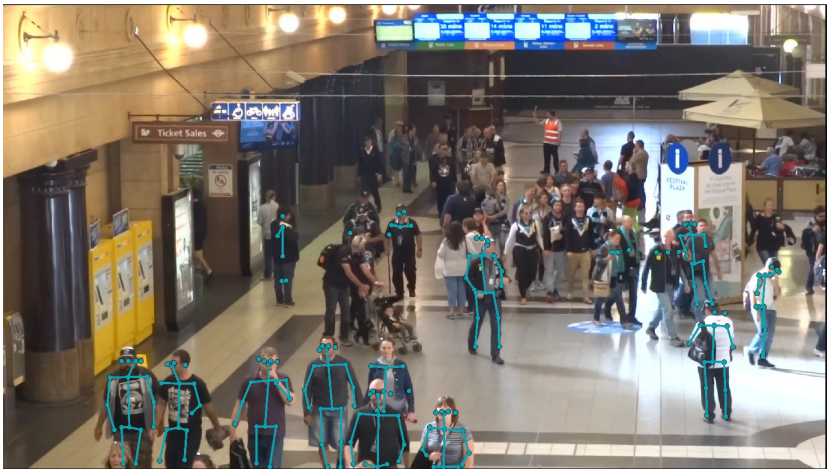
3.3.2 Transfer and Curriculum Learning

The annotation entity may use pre-annotation tools to generate annotation proposals. However, since at the beginning of the annotation process none or only scarce annotations are available, the tool requires prior training on similar data. If annotations are available, the tool can be fine-tuned on the target domain using transfer learning. In the case when no annotations are available, the tool can be used to generate a prliminary annotation proposal which then will be reviewed manually and progressively be retrained on the new available data. For instance, we consider the task of annotating the pose of all pedestrians in the dataset illustrated in Figure 2.1. At first, the annotation proposal tool has only been trained on COCO, an other target domain shown in Figure 3.3(a), where the ratio pixels per person is much higher. As expected, the results in Figure 3.4(b) and Figure 3.4(c) are acceptable on the first row. However, the persons behind them aren't recognized at all. After some frames have been manually annotated the pre-annotation tool can be trained using these annotations to create better key-point predictions. Ideally images of a lower difficulty level are first annotated and used for training. The difficulty is then progressively increased in accordance with the performance limits of the actual

(a) Image for annotation



(b) Keypoints pre-annotation results



(c) Keypoints pre-annotation results as skeleton representation

Figure 3.4: Keypoints pre-Annotation on MOT20 [5] using a pre-annotation tool trained on COCO [13]. Only the first row is detected which corresponds to the training domain.

instance of the tool. After a few iterations, this curriculum learning approach greatly reduces the effort for long and / or similar sequences.

3.3.3 Online Learning

Theoretically, considering the annotation process as the training phase of a model. Annotation proposals would represent a training iteration. The manually

corrected annotation is thus the ground truth. Therefore, the correction itself, e.g. the distance between the location of a proposed bounding box and the corrected one can be used to calculate an error which in turns may be used to update the parameter of the model. Consequently the resulting online learning enables immediate performance increase, without requiring complete retraining on the model and a long wait before the new deployment of the tool. Furthermore, the learned model should already conform to the annotation requirements.

3.3.4 Continuous Learning

Considering long time annotation projects with potentially continuous flows of new data to be annotated, different strategies may be required. Given a satisfactory annotation proposal tool trained with curriculum learning and / or online learning, we define an annotated subset, which is manually approved, as a validation set. The tool automatically annotates new incoming data and is periodically retrained over the ever growing dataset. The non-changing validation subset is then used to regularly assess the performance of the tool and thus prevent catastrophic forgetting or a negative feedback loop.

3.4 Discussion

A reliable, efficient and therefore highly automated annotation process is a complex and difficult process to model with need of constant improvement. Several interests and requirements are to be acknowledged and dealt with. Nevertheless, we identified great sources of improvement which can be addressed with efficient automation covering several aspects of machine learning which are currently topics of active and ever evolving research.

4 Data Quality Assurance

The annotation of data is a long, complex and repetitive process which requires intense concentration. Multiple potential issues may appear during annotation. The annotation may be incomplete, e.g. missing frames in a sequence or classes

of objects, or annotations in a wrong format, e.g. the pose has been annotated with 14 keypoints, but 17 were required. Different mistakes can be made such as misclassification, false positive, false negative or incorrect re-idenfication. Furthermore, annotations may lack precision, e.g. a keypoint lies a few pixels besides the intended point, a bounding box excludes the feet of a person. Throughout a longer sequence annotations maybe inconsistent, e.g. bounding box around the full body of a person instead of placing the bounding box around the visible body. Nevertheless, such issues are reliably detectable and therefore it is possible to address them within a short time. The data validation entity is responsible for methodically finding and reporting these issues. Following the annotation specifications, they perform a quality check on the annotation delivered by the data annotation entity. They provide detailed feedback on possible annotation issues to the data annotation entity and raise specification issues immediately. Specialized tools for review are used, thus the data validation entity itself is not permitted to perform the correction of the annotations.

In the remainder of this section the steps of the data quality assurance process is shortly reviewed, then the final step of data selection and validation is described.

4.1 Steps of the Data Quality Assurance Process

The steps of the data quality assurance process, as illustrated in Figure 4.1, are principally identical to the annotation steps. First the data validation entity analyzes the annotation specification and may use automation for annotation pre-checking. Then the validation task is distributed along the available personnel, performed and finally merged. Lastly the feedback and assigned issues, the validation artifacts, are reported to the data owner entity.

4.2 Data Selection and Validation

After ensuring sufficient quality of annotations, the dataset is split into three parts (with a possible fourth part of non assigned data): a training, validation and test subset.

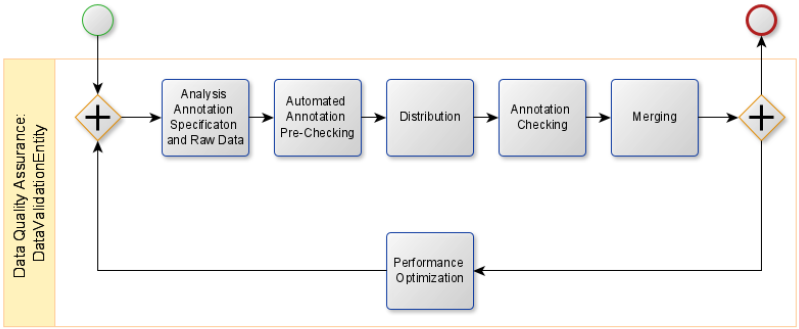


Figure 4.1: Data Quality Assurance Process Diagram

4.2.1 Training Subset

This subset is used to train the learned parameters for a supervised model. It should be carefully selected according to target use case and the scenario specifications, e.g. edge cases, ethic, law. The dataset should be representative, hence severe imbalance may result in discarding part of the annotated data.

4.2.2 Validation Subset

The validation subset is distinct from the training data. It is used to measure the performance of models during prototyping and training. It should reflect edge cases and offer sufficient diversity in regard to the target use case.

4.2.3 Test Subset

The test subset is used to measure the performance of the model after training in order to detect potential overfitting against the validation data. Therefore, the test subset is distinct from the training and validation subsets. Furthermore, it is used to asses the performance of a model against the target use case.

Each subset should be subject to review from the data curation entity against the target use case and scenario specifications. Finally, the whole process, intents and specifications should be rigorously documented to facilitate intern and extern audits, e.g. model cards [16] and datasheets [9].

5 Conclusion

Neural network aided data pre-annotation is a very promising approach. Yet not new. It is well-known for person detection or pose estimation in surveillance scenarios that specific problems such as false classification, person occlusion, split or merged detection often result from automatic predictions. A reliable and partly neural network aided annotation process should technically profit from human intelligence through specialized human operators and improve their abilities. Several methods were presented and discussed in this report with the focus on human activity in public place surveillance. Furthermore, concepts were presented for a reliable and transparent data annotation process, which naturally includes computer assisted steps through tailored neural networks. Future work will include the implementation and evaluation of these methods and also further investigations on a reliable and efficient annotation process for scenes including multiple cameras oriented on one and a same place.

References

[1] Joy Buolamwini and Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Conference on fairness, accountability and transparency*. 2018, pp. 77–91.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. “COCO-Stuff: Thing and stuff classes in context”. In: *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE. 2018.

[3] *Computer Vision Annotation Tool (CVAT)*. `https://github.com/openvinotoolkit/cvat/`.

[4] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[5] P. Dendorfer et al. “MOT20: A benchmark for multi object tracking in crowded scenes”. In: *arXiv:2003.09003[cs]* (Mar. 2020). arXiv: 2003.09003. URL: http://arxiv.org/abs/1906.04567.

[6] J. Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR09*. 2009.

[7] Matthis Eicher et al. *Whitepaper Process considerations: A reliable AI data labeling process*. Aug. 2020. URL: https://wiki.eclipse.org/images/0/0e/WhitePaper_Process_considerations-A_reliable_AI_data_labeling_process.pdf.

[8] M. Everingham et al. “The Pascal Visual Object Classes (VOC) Challenge”. In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.

[9] Timnit Gebru et al. “Datasheets for datasets”. In: *arXiv:1803.09010* (2018).

[10] Andreas Geiger et al. “Vision meets Robotics: The KITTI Dataset”. In: *International Journal of Robotics Research (IJRR)* (2013).

[11] HEG-KI. *Ethics guidelines for trustworthy AI*. 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[12] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. “Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning”. In: *Advances in neural information processing systems*. 2019, pp. 7026–7037.

[13] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[14] John Markoff. *Seeking a Better Way to Find Web Images*. Nov. 2012. URL: https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html.

[15] Margaret Mitchell et al. “Diversity and inclusion metrics in subset selection”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 117–123.

[16] Margaret Mitchell et al. “Model cards for model reporting”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[17] Gerhard Neuhold et al. “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes”. In: *International Conference on Computer Vision (ICCV)*. 2017. URL: https://www.mapillary.com/dataset/vistas.

[18] Inioluwa Deborah Raji et al. “Saving face: Investigating the ethical concerns of facial recognition auditing”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 145–151.

[19] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[20] Daniel Rödler. *A Feast of Sensor Data: Feeding Self-driving Algorithms*. Nov. 2020. URL: https://understand.ai/blog/annotation/autonomous-driving/machine-learning/2020/11/12/feast-of-sensor-data-feeding-self-driving-algorithm.html.

[21] Ozan Sener and Silvio Savarese. “Active Learning for Convolutional Neural Networks: A Core-Set Approach”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=H1aIuk-RW.

[22] *Sloth*. https://github.com/cvhciKIT/sloth/.

[23] Fedor Zhdanov. *Diverse mini-batch Active Learning*. 2019. arXiv: 1901.05954 [cs.LG].

Improving 3D Semantic Segmentation with Twin-Representation Networks

Fabian Duerr

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
fabian.duerr@audi.de

Abstract

The growing importance of 3d scene understanding and interpretation is inherently connected to the rise of autonomous driving and robotics. Semantic segmentation of 3d point clouds is a key enabler for this task, providing geometric information enhanced with semantics. To use Convolutional Neural Networks, a proper representation of the point clouds must be chosen. Various representations have been proposed, with different advantages and disadvantages. In this work, we present a twin-representation architecture, which is composed of a 3d point-based and a 2d range image branch, to efficiently extract and refine point-wise features, supported by strong context information. Additionally, a feature propagation strategy is proposed to connect both branches. The approach is evaluated on the challenging SemanticKITTI dataset [2] and considerably outperforms the baseline overall as well as for every individual class. Especially the predictions for distant points are significantly improved.

1 Introduction

Understanding a 3d environment is one of the key challenges for autonomous vehicles or robots. For this task of 3d scene understanding and interpretation,

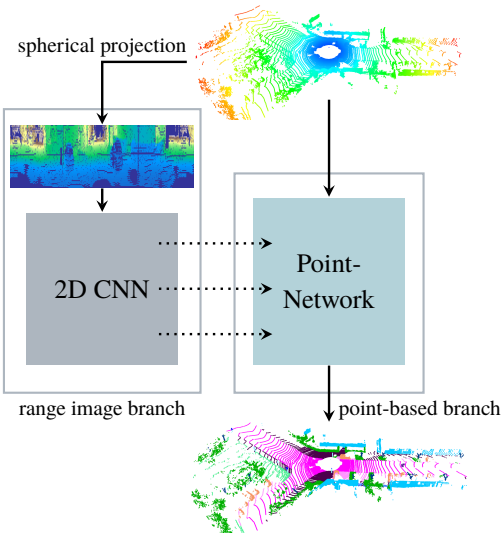


Figure 1.1: The proposed twin-representation architecture, which exploits two different point cloud representations. The 3d point-based branch extracts and refines point-wise features while the 2d range image branch efficiently aggregates context information.

semantic segmentation of images or point clouds, which assigns a class label to every pixel or 3d point, provides valuable information.
The combination of geometric and semantic information provided by 3d semantic segmentation is particularly valuable. To tackle this task with established deep learning approaches, like Convolutional Neural Networks (CNNs), a proper representation of point clouds has to be chosen, to allow their application. Point-based approaches [12, 19] operate directly on the raw point clouds while projection based methods [11, 18] transform them into a regular space, like 2d or 3d grid, to enable convolution operations.
Recently, the combination of voxel and point-based representation showed promising results [9, 17], by exploiting the advantages of both representations. In general, projection based methods, like voxel grids, efficiently aggregate neighborhood information because of the regularity of their data representation. The projection however requires a discretization in most cases, where the choice

of resolution is a trade-off between loss of information and memory as well as computational costs. Point-based approaches on the other hand efficiently operate on the original point cloud resolution without information loss, but the aggregation of neighborhood information and context is expensive. Because of these complementary properties, the combination of both representations offers a great potential.
This work follows the general idea of combining projection and point-based representations but focuses on the more efficient range image representation. Therefore, we present a twin-representation architecture, which combines a 2d range image and a 3d point-based branch, see Fig. 1.1. The 2d branch works on range images resulting from a spherical projection and enables the efficient aggregation of local neighborhoods and context. The point-based branch computes point-wise features while preserving the original resolution and is supported by the aggregated information from the 2d branch, to predict the final 3d semantic segmentation. To summarize, our contributions are twofold:

- A twin-representation architecture composed of a 2d range image and 3d point branch, which preserves point-wise features while efficiently aggregating local context.
- A feature propagation strategy for 2d $\rightarrow$ 3d feature transformation.

2 Related work

The growing importance of autonomous vehicles and robots also raised the importance of 3d semantic segmentation. Supported by an increasing number of available indoor [1] and outdoor datasets [2, 23, 3] considerable progress has recently been achieved. A crucial and recurring question when addressing 3d semantic segmentation with CNNs is the representation of 3d point clouds. Many different representations have been proposed in recent works, which can generally be grouped into two categories.
Point-based methods, like PointNet [12] and its successor PointNet++ [13], directly process the raw point clouds. PointNet applies a shared multilayer perceptron (MLP) pointwise and a symmetric operation performs global feature

aggregation. While this is very efficient, a single global feature aggregation greatly limits the ability to capture spatial relations. Therefore, PointNet++ was proposed, which applies individual PointNets to local regions and aggregates them in a hierarchical fashion. While being one of the first approaches, many others [8, 7, 20, 19] followed.
Projection based methods can further be divided into subcategories based on the chosen regular space, like voxel grids [27, 18], permutohedral lattice [15, 16] or bird's eye view [26]. Another possibility is a spherical projection, which results in a so called range image. SqueezeSeg [21] was one of the first approaches building upon range images for a road segmentation task. Improved versions were released in [22] and [24]. The latter targets full semantic segmentation and proposed Spatially-Adaptive Convolutions (SAC) to deal with spatially-varying feature distributions, induced by the spherical projection. Another approach is RangeNet++ [11], which builds upon the DarkNet53 backbone [14] and presented a label projection strategy from range image space to 3d point clouds. [10] proposed LaserNet, based on deep layer aggregation [25], for 3d object detection, while one intermediate result is a semantic segmentation.
Recently, first attempts were made to exploit the advantages of multiple representations in one architecture. PVCNN [9] combined a shared MLP for point-wise feature extraction with 3d convolutions in voxel space for context aggregation. It is therefore able to extract point features in full resolution while extracting and aggregating neighborhood information in a coarse voxel space. It's successor SPVCNN [17] replaced the dense 3d convolutions by its spare counterparts, which allows for a higher voxel resolution and therefore more preserved information. While sparse 3d convolutions already improve the performance and possible resolution, 2d convolutions are still more efficient with similar or less information loss. Therefore, our proposed segmentation architecture combines a 2d range image branch with a point-based branch and relies on a novel feature propagation strategy from 2d range image space to 3d point clouds.

3 Twin-Representation Network

The goal of the presented approach is the exploitation of two different input representations, range images and 3d point clouds, to improve 3d semantic segmentation. Fig. 3.1 shows the overall architecture, which consists of three main components. The range image backbone provides 2d feature maps of different stages and resolution while a feature propagation step transforms the 2d features back to their corresponding 3d points. Thereby, both components together efficiently provide aggregated neighborhood and context information for each individual point. These are used by the third component, a 3d point network. In the following, we provide details for each individual component.

Range Image Backbone Range images and the corresponding spherical projection are motivated by a lidar's internal structure, which usually consists of a vertical stack of lasers spinning around their vertical axis. As a result, the measurements can be described by an azimuth angle ϕ, an elevation angle θ and measured distance r and intensity e. We follow [4] for the conversion of the point clouds to range images of shape $6 \times h \times w$, with channels r, x, y, z, e and an occupancy flag.
The chosen 2d network architecture is based on deep layer aggregation [25] and closely related to LaserNet [10]. We reduced the number of Residual Units [6] in the first two feature extractors to four and five. Additionally, the downsampling in the first feature extractor was omitted. The backbone provides 2d feature maps of three different stages, see Fig. 3.1. Because of the underlying deep layer aggregation, all three stages are at full resolution while still representing features of different context stages. Full resolution feature maps have the advantage, that a distinctive feature vector can be provided for every 3d point, expect for colliding points [4].

Feature Propagation The fusion of feature maps $\boldsymbol{F}$ and point features $\boldsymbol{f}^{\text{point}}$ requires a transformation of 2d features back to their corresponding 3d points $\boldsymbol{p}$. One possible strategy is the assignment of a 2d feature to the 3d point belonging

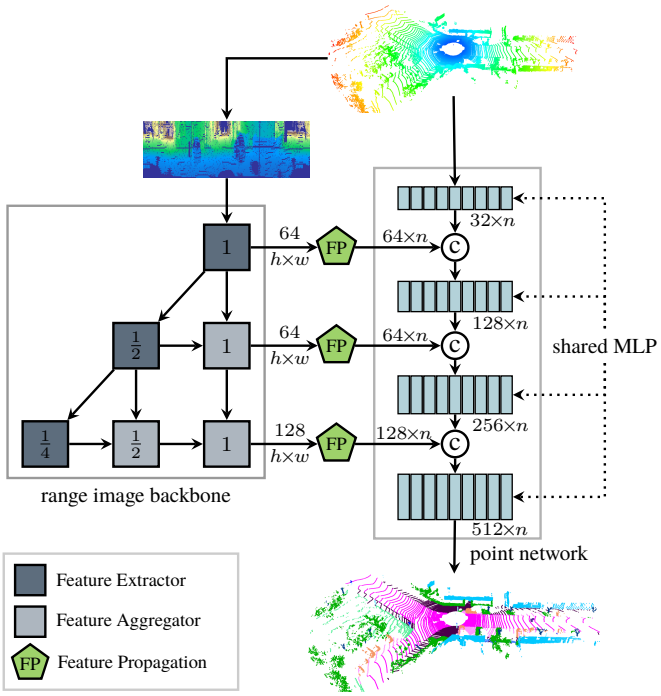


Figure 3.1: The proposed segmentation architecture. The range image backbone efficiently aggregates 2d context information and provides them via a feature propagation step to the point network, which itself computes point-wise features and combines them with the provided context information.

to its pixel position

$$\boldsymbol{f}_{\boldsymbol{p}}^{\text{twin}} = \boldsymbol{f}_{\boldsymbol{p}}^{\text{point}} \oplus \boldsymbol{F}[u_{\boldsymbol{p}}, v_{\boldsymbol{p}}], \tag{3.1}$$

where u and v are the 2d coordinates resulting from the spherical projection and $\oplus$ denotes concatenation. One possible disadvantage of this strategy occurs for colliding points, because all of them get the same feature vector assigned, even if they are far apart in 3d. For solving the related challenge of label back-projection, [11] proposed a KNN-based approach. The 3d labels are chosen by a majority vote among the approximated k-nearest-neighbors, weighted by their euclidean

distance. Although we want to back-project intermediate feature vectors, instead of simple labels, we yet pick up the general idea of using the 2d neighborhood of a pixel as nearest-neighbor candidates. Instead of a majority vote, we compute the weighted sum over the feature vectors:

$$\boldsymbol{f}_{\boldsymbol{p}}^{\text{twin}} = \boldsymbol{f}_{\boldsymbol{p}}^{\text{point}} \oplus \sum_{\tilde{\boldsymbol{p}} \in \mathcal{N}_k(\boldsymbol{p})} w_{(\boldsymbol{p},\tilde{\boldsymbol{p}})} \cdot \boldsymbol{F}[u_{\tilde{\boldsymbol{p}}}, v_{\tilde{\boldsymbol{p}}}], \quad w_{(\boldsymbol{p},\tilde{\boldsymbol{p}})} = \frac{1}{||\boldsymbol{p} - \tilde{\boldsymbol{p}}||^2}, \quad (3.2)$$

with $\mathcal{N}_k$ being the $k \times k$-neighborhood of $\boldsymbol{p}$. Therefore, features are aggregated based on the point distribution in 3d space.

Point Network Motivated by the original PointNet, the point network stacks multiple shared MLPs to extract and refine point features. After each stage, the propagated features from the 2d branch, which provide the aggregated neighborhood and context information, are concatenated with the point features, see Eq. 3.1 and 3.2. The point network operates on the original point cloud resolution over all stages, so no information are lost. The shared MLPs are implemented by 1×1-convolutions and their feature channel depth increases with network depth.

4 Experiments

4.1 SemanticKITTI

We evaluate our approach on the challenging, large-scale SemanticKITTI dataset [2, 5], which provides point-wise annotations for $360°$-Velodyne-HDL-64E scans. The annotations contain 19 classes for the single scan benchmark. 22 labeled sequences of varying length, recorded at $10\,\text{Hz}$, add up to just over $43,000$ scans. Sequences 0-10 are provided with labels for training and validation while sequences 11-21 without published labels form the test split. The official recommendation is to use sequence 08 for validation, but we use a larger validation split for our ablation studies, consisting of sequences $02, 06, 10$, for more significant conclusions. We follow the official evaluation metric and report the mean Intersection-over-Union (mIoU).

Table 4.1: Observed improvements when adding the point network (PN) and KNN feature propagation, compared to a single range image backbone (RB).

RB	PN	KNN	mIoU (%)
✓			51.4
✓	✓		53.7
✓	✓	✓	**54.8**

4.2 Implementation Details

The implementation is based on PyTorch and all experiments are trained in mixed precision mode using distributed data parallel training on four Tesla V100 GPUs.
Class-balanced cross entropy loss is optimized by Adam with a weight decay of 0.0005 for $100k$ iterations. The learning rate starts with 0.001 and is then multiplied with $e^{-5 \cdot 10^{-5} \cdot i}$ after every iteration i. To counteract overfitting, we randomly flip the range images horizontally with a probability of $p = 0.5$ and rely on random crops of size 64×1024 during training.
First, solely the range image backbone is trained with a batch size of 32. Building upon this, we train the entire network, also with a batch size of 32.

4.3 Results

Our evaluation starts with an investigation of the influence of the individual components, with the results being depicted in Table 4.1. The range image backbone, as a common 2d range image approach, is our baseline and achieves a mIoU of 51.4%. The presented twin-representation architecture, which is composed of the backbone and a point network, significantly outperforms the baseline by $+2.3\%$. Replacing the simple propagation strategy by the proposed KNN feature propagation further improves the results to 54.8%.
In the next step, we investigate the results restricted to the distance intervals $0-20$m, $20-40$m and >40m. Table 4.2 shows an overall performance increase of our approach for all chosen intervals. However, especially the results for distant

Table 4.2: Comparison of the mIoU (%) for different distance intervals.

Approach	0−20m	20−40m	>40m
RB	52.7	43.6	33.9
RB+PN	54.9	46.7	36.7
RB+PN+KNN	**55.6**	**47.2**	**39.2**

points are significantly improved by $+5.3\%$, which is particularly challenging because of the declining point density with increasing distance. For the other two intervals, a smaller but still considerable improvement of $+2.9\%$ and $+3.6\%$ is achieved.
Finally, we evaluate the results for the individual classes. Looking at Table 4.3, especially the classes motorcycle, truck, person and bicyclist experience a significant improvement by using the combination of range image backbone and point network. Likewise, the results for the classes car, other-vehicle, trunk and pole improved. In general, while no significant improvements for greater static classes can be observed, small classes greatly benefit from our approach. Adding KNN feature propagation further improves the results for most classes, without any bias regarding a special group of classes. One class to emphasize however is motorcyclist, which is improved by $+11.2\%$.

5 Conclusion

In this work, we presented a twin-representation architecture to combine a 3d point-based branch with a 2d range image branch, to improve 3d semantic segmentation. While the first computes and refines point-wise features over multiple stages, the latter supports the 3d branch with an efficient aggregation of neighborhood and context information. A feature propagation step connects both branches. The evaluation showed a significant overall improvement, considering that our approach outperforms the baseline for every individual class. Additionally, especially distant points experience a significant improvement. To summarize, combining the two input representations enables the exploita-

Table 4.3: Overview over the improvements for the individual classes. The presented approach outperforms the baseline for every single class. Values are given as IoU (%).

Approach	mIoU	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist
RB	51.4	83.9	31.7	35.9	33.4	31.1	45.4	23.2	2.4
RB+PN	53.7	85.0	**32.3**	44.1	42.5	34.7	**53.6**	**29.0**	3.1
RB+PN+KNN	**54.8**	**86.5**	29.7	**45.9**	**44.4**	**36.2**	53.0	28.4	**14.3**

Approach	road	sidewalk	parking	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic sign
RB	91.2	80.2	58.8	8.6	76.7	58.0	82.8	63.8	70.1	49.5	49.6
RB+PN	90.8	80.0	59.0	**8.8**	76.5	58.3	82.9	**66.2**	70.9	**52.0**	50.4
RB+PN+KNN	**91.3**	**80.5**	**61.0**	7.3	**78.1**	**60.2**	**83.4**	65.6	**71.7**	49.3	**52.8**

tion of their different strengths, which considerably improves 3d semantic segmentation.

References

[1] Iro Armeni et al. “3D Semantic Parsing of Large-Scale Indoor Spaces”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[2] Jens Behley et al. “SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2019.

[3] Holger Caesar et al. “nuScenes: A multimodal dataset for autonomous driving”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[4] Fabian Duerr et al. “Iterative Deep Fusion for 3D Semantic Segmentation”. In: *IEEE International Conference on Robotic Computing (IRC)*. 2020.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? The KITTI vision benchmark suite”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[6] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[7] Qingyong Hu et al. “RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[8] Yangyan Li et al. “PointCNN: Convolution On $\mathcal{X}$-Transformed Points”. In: *Advances in Neural Information Processing Systems*. 2018.

[9] Zhijian Liu et al. “Point-Voxel CNN for Efficient 3D Deep Learning”. In: *Advances in Neural Information Processing Systems*. 2019.

[10] Gregory P. Meyer et al. “LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[11] A. Milioto et al. “RangeNet++: Fast and Accurate LiDAR Semantic Segmentation”. In: *IEEE International Conference on Intelligent Robots and Systems (IROS)*. 2019.

[12] Charles Ruizhongtai Qi et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[13] Charles Ruizhongtai Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *Advances in Neural Information Processing Systems*. 2017.

[14] Joseph Redmon and Ali Farhadi. “YOLOv3: An Incremental Improvement”. In: *ArXiv*. Vol. abs/1804.02767. 2018.

[15] Radu Alexandru Rosu et al. “LatticeNet: Fast Point Cloud Segmentation Using Permutohedral Lattices”. In: *Robotics Science and Systems (RSS)*. 2020.

[16] Hang Su et al. “SPLATNet: Sparse Lattice Networks for Point Cloud Processing”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[17] Haotian Tang et al. “Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution”. In: *IEEE European Conference on Computer Vision (ICCV)*. 2020.

[18] Lyne P. Tchapmi et al. “SEGCloud: Semantic Segmentation of 3D Point Clouds”. In: *International Conference on 3D Vision (3DV)*. 2017.

[19] Hugues Thomas et al. “KPConv: Flexible and Deformable Convolution for Point Clouds”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2019.

[20] Shenlong Wang et al. “Deep Parametric Continuous Convolutional Neural Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[21] Bichen Wu et al. “SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2017.

[22] Bichen Wu et al. “SqueezeSegV2: Improved Model Structure and Unsupervised Domain Adaptation for Road-Object Segmentation from a LiDAR Point Cloud”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2018.

[23] Jun Xie et al. “Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[24] Chenfeng Xu et al. “SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[25] Fisher Yu et al. “Deep Layer Aggregation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[26] Chris Zhang, Wenjie Luo, and Raquel Urtasun. “Efficient Convolutions for Real-Time Semantic Segmentation of 3D Point Clouds”. In: *International Conference on 3D Vision (3DV)*. 2018.

[27] Yang Zhang et al. “PolarNet: An Improved Grid Representation for Online LiDAR Point Clouds Semantic Segmentation”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

"Let's get ready to bundle!": Crowd-level Human Keypoint Tracking

Thomas Golda

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

Abstract

This work examines the suitability of a state-of-the-art human pose tracking method for application within surveillance scenarios and focuses on public places in urban areas that tend to suffer from crowdedness, such as city centers. Starting with a short introduction to motivate keypoint tracking in surveillance applications, this report will present details about the adapted method, which follows an LSTM-based approach. Afterwards, different changes that had to be incorporated in order to successfully apply the given method to our target setting will be presented. Finally, various experiments will show how the chosen method performs, based on experiments with simulated data.

1 Introduction

Anomaly detection amongst other strongly related topics like outlier and novelty detection, plays an important role in various research fields as network traffic monitoring, time series analysis, medical image analysis, and video surveillance. However, when talking about anomalies in the context of video surveillance the understanding of what an anomaly actually looks like can differ strongly

between applications. For instance, an anomaly can be an abandoned suitcase at a public place, a vehicle driving through a pedestrian zone or suspicious or salient behaving people. With the rising interest in unconstrained activity and action recognition[1] in urban settings and application-oriented research for video surveillance, the task of detecting unusual behavior gets more and more attention. This report focuses on human-centered features in order to distinguish between usual and unusual behavior. Therefore, on a basis of person skeletons provided by human pose estimators, we try to create corresponding body joint tracklets, as proposed in [3]. In order to do so, different keypoints corresponding to a certain person have to be tracked over time to obtain the desired tracklets, which can afterwards be used for further behavioral analysis.

2 Tracking Keypoints in the Wild

2.1 Human Pose Estimation

Human Pose Estimation describes the problem of estimating a skeletal representation of a person based on information gathered using certain types of sensors. The skeletal representation is typically represented as a graph $G = (V, E)$ where $V \subset \mathbb{R}^n$ is a set of keypoints and $E \subset V \times V$ is a set of edges connecting various keypoints. Depending on the chosen skeletal model the graph can be seen as a tree. In general, Human Pose Estimation considers sensors used in classical video cameras or depth cameras delivering RGB or RGB-D information respectively. However, in the field of video surveillance RGB-D cameras are rarely applied, which is due to price and often large distances of subjects to the mounted camera. Therefore, this work focuses on the case of 2D skeletons obtained using classical cameras and RGB data. With this constraint, the resulting skeletons produced by human pose estimation algorithms consequently consist of keypoints in a two-dimensional space with $V \subset \mathbb{R}^2$.

[1] https://actev.nist.gov/

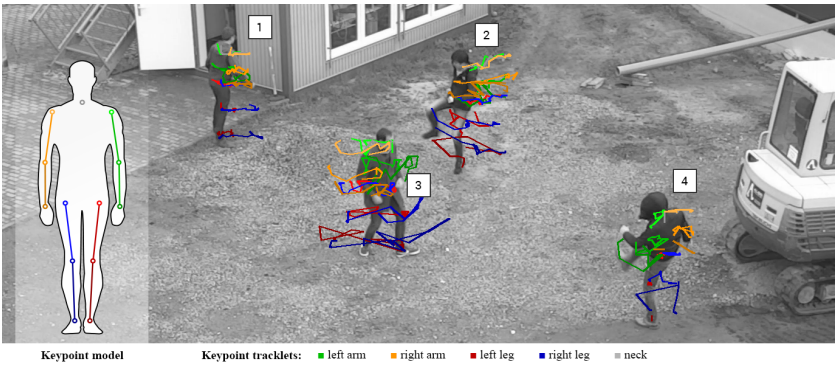


Figure 2.1: Crop of exemplary network camera footage. This is an example of keypoint bundles extracted using a simple tracking-by-detection approach. Based on detections in t consecutive frames, for every person keypoint tracklets were extracted. The different resulting body parts of the keypoint bundles are highlighted in different colors for visualization purposes. In addition to that, the lightness of those colors encodes the proximity of the corresponding keypoint to the pose model center, i.e. shoulders and hips have lighter colors than wrists and ankles. For a better understanding, the underlying human pose model with the corresponding colors is shown on the left.

2.2 Human Keypoint Tracking

2.2.1 Introduction

In general, the process of observing a single entity over time is referred to as *tracking*. Classical tasks are object and pedestrian tracking, where for a given timestep detections of single entities are associated with corresponding detections in earlier timesteps. This kind of tracking is also known as tracking-by-detection. Typically, such entities are represented by bounding boxes that enclose the subject. The idea of tracking bounding boxes can be extended to the tracking of single points over time, which is subject of this report.
Given a video or image sequence showing pedestrians our goal is to obtain a set of tracklets K_i for every single person i describing its movement over a short period of time. The set of tracklets is defined as

$$K_i = \{(\mathbf{k}_1^1, ..., \mathbf{k}_t^1), (\mathbf{k}_1^2, ..., \mathbf{k}_t^2), ..., (\mathbf{k}_1^n, ..., \mathbf{k}_t^n)\} \tag{2.1}$$

where $\mathbf{k}_t^n \in \mathbb{R}^2$ is the n-th two-dimensional keypoint at timestep t. Hence, each item in K_i is a single corresponding keypoint tracklet. We refer to K_i as a *keypoint bundle*. Such keypoint bundles can be obtained using various approaches. In this work, we assemble these bundles by tracking keypoints directly in order to create less noisy tracklets than those acquired by following an approach just based on person detection. Figure 2.1 shows an example for a scene and corresponding keypoint bundles.

2.2.2 Multi-person Pose Tracking using Sequential Monte Carlo with Probabilistic Neural Pose Predictor

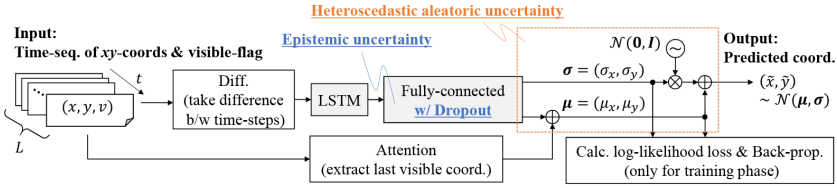


Figure 2.2: The figure shows the structure of the probabilistic Neural Pose Predictor by Okada et al. [6]. Whereas the epistemic uncertainty is modelled using dropout in the final fully-connected layer of the LSTM-architecture, the aleatoric uncertainty is incorporated using an sampling process from a normal distribution.

Okada et al. [6] followed the approach of tracking keypoints by using an LSTM-based network architecture called *(probabilistic) Neural Pose Predictor* (NPP). In general, the NPP is part of a classical particle filter where it is responsible for modeling the motion model of humans and hence is an essential part of the prediction step within the update process of the bayesian filter. The main contribution of Okada et al. is the incorporation of different uncertainties within the NPP, namely heteroscedastic aleatoric uncertainty (inherent system stochasticity) and epistemic uncertainty (model uncertainty due to limited data). Based on a fixed time horizon of ten timesteps the position at the next time step is predicted, which is done for every single keypoint and every detected pedestrian. To be exact, input to the NPP is a sequence of differences between consequent timesteps in a single track. The LSTM then predicts values for the

mean and standard deviation, which are used to define a normal distribution that is afterwards used to sample an actual estimate for the predicted difference. The explained procedure is also shown in Figure 2.2.

3 From One to Many

We adapted the approach of Okada et al. [6] in order to evaluate its performance on crowd-level applications, which is a topic rarely covered in the computer vision community. It is obvious that this is a very challenging task, due to various problems like the ambiguity in appearance and lots of dynamic occlusions. Figure 3.1 shows how such a crowded surveillance situation can look like. Although pedestrians in this image are at a quite large scale, already the proximity to other pedestrians can lead pose estimators to estimate false poses. This is supported by many people wearing dark clothes, which due to less

Figure 3.1: Example of a crowded scene. People in such monitored areas are typically even of smaller size, however independent of the scale the same problems occur: ambiguity in appearance and strongly dynamic occlusions.

contrast is very challenging for pose estimation methods. However, tackling it in this way offers more detailed information about the behavior of single subjects and still keeps a high level of data privacy at the same time. This is especially useful when it is used in assistance systems for video surveillance, where the system just creates hints for a human operator. In our experiments we therefore are mainly interested in the overall keypoint tracking performance.

3.1 SyMPose

Especially designed for being a dataset consisting of many people and providing detailed keypoint and tracking information for each individual, we decided to use an internal synthetic dataset for for training the NPP. This dataset was created using an own optimized version of the JTA modification published by Fabbri et al. [2] for the popular video game *Grand Theft Auto V*. It was initially designed for the task of domain adaption between synthetic and real-world domain [4]. Table 3.1 gives a short comparison of some key figures of both datasets.

Table 3.1: Comparison of two synthetic datasets. Although SyMPose is smaller compared to JTA, it is more comparable to a surveillance situation due to the higher viewing position of the camera and the overall higher average pedestrian density (ppf) within a frame. Furthermore, the maximum number of pedestrians per frame is more than twice as high than in JTA (130 versus 60 pedestrians).

dataset	# scenes	# frames	# poses	ppf	setting
JTA [2]	512	460,800	ca. 10m	21	urban
SyMPose [4]	21	19,900	ca. 1.3m	68	urban

In order to get a better idea of how scenes from SyMPose look like, an example taken from the dataset is displayed in Figure 3.2. Most noticeable is the much higher number of pedestrians in the scene. Furthermore, the viewing perspective and camera mounting height are more consistent throughout all recorded scenes and can be hence better compared to a real-world application scenario. These were the main reasons to use this dataset, since comparable real-world datasets labeled with keypoint information in this setting do not exist or are publicly just not available. SyMPose was used throughout this work for training and evaluation.

Figure 3.2: Synthetic crowded scene. SyMPose consists of around two dozen scenes showing different places filled with many people. All scenes are recorded from a higher-level camera that should simulate the target scenario in urban settings. The underlying skeleton model consists of a subset of the JTA model and was designed to mimic the model used in CrowdPose [5].

3.2 Whole Pose Inference

In the initial proposed method, Okada et al. [6] designed the NPP to work on single keypoints. This, however, includes the assumption that every keypoint has the identical, independent motion model. Apparently, the human skeleton is restricted in its configurations and the movement of a single joint (i.e. keypoint) is strongly dependent on its adjacent joints. We therefore extended the single-keypoint approach to a multi-keyoint approach and compare it with the former. This is done by providing a feature vector to the NPP consisting of information for all keypoints instead just for a single one.

3.3 Expansion of Inter-Frame Distances

Since recorded data in such surveillance scenarios is dominated by people at small scale and slow velocities, the relative movement between two frames is

also quite small compared to a close-up shot of only few people. With a typical recording frame rate of 25 or 30 frames per second this results in distances of only few pixels between two poses or frames. This is a challenging problem for the NPP, which was designed to work with multi-person scenarios yet at larger scale.

3.3.1 Affine Transformation

A first way to tackle this problem is to rescale tracks by a certain factor $\alpha \in \mathbb{R}^+$. Typically, such a value is determined by normalizing the data using its mean and standard deviation. Doing so results in an affine transformation and will affect the size of the poses and hence the position of the keypoints as well as the displacement between timesteps. The benefit of this approach is, that still the full frame rate and therefore the full range of information can be used. However, this transformation comes with additional lightweight computations and has to be reverted afterwards to extract the real coordinates of the predicted position.

3.3.2 Reduction of Frame Rate

Another orthogonal way to tackle this problem is to skip some intermediate frames, and hence sample a given sequence. This will have a similar effect as the first way, however it is associated with a loss of information since fast movements could be overlooked. Furthermore, the prediction frame rate is also affected by the new resulting frame rate. The benefit of this approach is that no transformation is needed and the predicted pose can be used directly.

4 Evaluation

In the following, we will examine the performance of the changes and extensions presented in Section 3. First the used evaluation metrics will shortly be presented and analyzed. These metrics were chosen, since they were also used by Okada et al. [6].

4.1 Evaluation Metrics

4.1.1 Mean Squared Error

The Mean Squared Error (MSE) is a widely used metric to measure the deviation from a certain target. In the context of human poses it is defined as

$$\text{MSE} = \frac{1}{2N}\sum_{i=1}^{N}(\hat{k}_{i,x} - k_{i,x})^2 + (\hat{k}_{i,y} - k_{i,y})^2 \tag{4.1}$$

where $N \in \mathbb{N}$ is the number of keypoints defined by the underlying pose skeleton model, $\hat{\mathbf{k}}_i = (\hat{k}_{i,x}, \hat{k}_{i,y})$ is the ground truth position for the i-th keypoint and $\mathbf{k}_i = (k_{i,x}, k_{i,y})$ is the position of the prediction of the corresponding keypoint. It is obvious, that the value of the MSE is strongly dependent on the scale of two compared poses.
To tackle this scale dependency, we us a normalized version of the MSE. By dividing the MSE by the squared scale factor α we obtain the following equation

$$\alpha\text{MSE} = \frac{\text{MSE}}{\alpha^2} = \frac{1}{2N\alpha^2}\sum_{i=1}^{N}(\hat{k}_{i,x} - k_{i,x})^2 + (\hat{k}_{i,y} - k_{i,y})^2 \tag{4.2}$$

4.1.2 Object Keypoint Similarity

The Object Keypoint Similarity (OKS) [1] is a metric introduced by the COCO consortium for the purpose of evaluating the pose detection performance. It is an attempt to create a metric that can be compared to the Intersection of Union, which is especially known for its application in bounding box detection scenarios.

$$\text{OKS} = \frac{\sum_i [exp(\frac{-d_i^2}{2 \cdot s^2 \kappa_i^2}) \cdot \delta(v_i > 0)]}{\sum_i [\delta(v_i > 0)]} \tag{4.3}$$

where N is the number of keypoints in the underlying skeleton model, s is the square root of the area of the smallest bounding box enclosing all corresponding keypoints and d is the actual distance between two poses. Furthermore, v_i is the

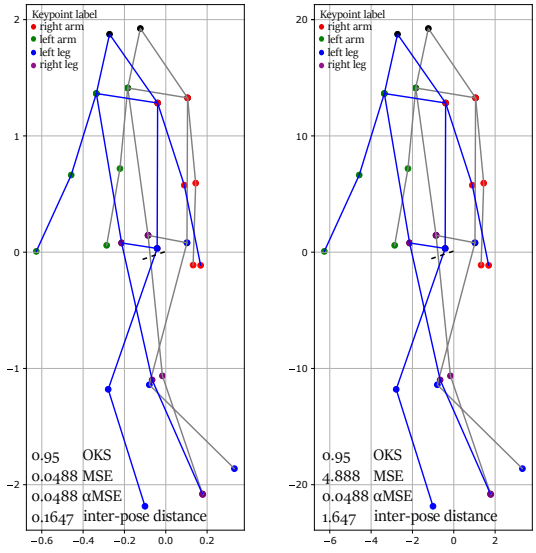


Figure 4.1: Visual and quantitative comparison for poses at different scales and the effects on the evaluation metrics OKS and MSE. Inter-pose distance and MSE are directly effected by the rescaling of poses, whereas OKS and αMSE stay the same due to their scale invariance.

visibility of the i-th keypoint and κ_i are manually obtained keypoint constants that model the annotation uncertainty of human annotators in the actual COCO dataset. Due to the way the OKS is defined, the metric is scale invariant. This comes from the fact that the actual keypoint positions are combined with the occupied area, which can be seen as a certain kind of normalization.

Figure 4.1 shows two examples of pose comparisons. Both examples look very similar, however they are at different scale. While the OKS stays the same even at larger scale, the MSE is higher the bigger the scale is.

4.2 Evaluation Results

First, we take a look on the presented changes and extensions. Table 4.2 summarizes the results for single- and multi-keypoint approach, as well as both inter-frame distance expansion ways. In order to compare MSE results between different scales, we report the αMSE introduced in Section 4.1.

Single- vs. Multi-Keypoint Throughout all examined configurations the single-keypoint approach achieves similar or even slightly better results compared to the multi-keypoint approach. With increasing scaling factor α both the single-keypoint and multi-keypoint approach improve with regard to the OKS and αMSE metric. Furthermore, with the the expanding absolute gap between timesteps, the single-keypoint approach builds up a lead over the multi-keypoint method. This, however underlines that using a NPP just trained on single keypoints might be sufficient and does not benefit from the additional information that is available when predicting whole poses.

Affine Transformation In order to evaluate the impact of the affine transformation, different values for α were chosen, namely $\alpha = 10^0$ (i.e. no scaling), $\alpha = 10^1$, $\alpha = 10^2$, $\alpha = 10^3$, and $\alpha = 10^4$. The data distribution is characterized by its mean μ and its standard deviation σ which are given in Table 4.1 for every examined frame rate at scale $\alpha = 1$. The values show that the normalization would yield a scaling factor α between roughly around 2 and 12, and hence would fall into the range of our experiments with $\alpha = 1$ and $\alpha = 10$. As mentioned above, independent from the chosen NPP-approach (single- vs. multi-keypoint), with increasing values of α the prediction performance first improves and drops for larger values of α. The results show, that choosing a larger rescaling factor than the factor induced by normalization is beneficial to the performance of the NPP and leads to significant improvement. The same observation can not only be made for the full, but also for the reduced frame rate experiments. One conceivable reason for this behaviour might be that the LSTM architecture of the NPP struggles with very small values close to zero. This is likely to happen since input and output of the LSTM are differences

Table 4.1: Mean and standard deviation for the distance between consequent poses. Since intermediate frames were dropped for the reduced experiments, the actual distance between consequent frames was increased throughout the dataset.

Framerate [fps]	**30**	**6**	**3**
μ	0.1036	0.3644	0.6367
σ	0.0802	0.2685	0.4892

Table 4.2: Evaluation results on the SyMPose test set. With decreasing number of frames the performance of the pose prediction drops. This is logical since less information about the actual motion is available. Furthermore, with an increasing factor α the prediction performance improves up to a value of around 10^2 to 10^3 and begins to drop afterwards. This result shows the weakness of LSTMs when used with small values.

Framerate [fps]		**30**		**6**		**3**	
	α	**OKS**	**αMSE**	**OKS**	**αMSE**	**OKS**	**αMSE**
single	10^0	0.989	0.02	0.75	0.23	0.58	0.73
multi		0.989	0.02	0.75	0.23	0.52	0.78
single	10^1	0.990	0.02	0.87	0.13	0.78	0.26
multi		0.990	0.02	0.86	0.13	0.79	0.28
single	10^2	0.996	**0.01**	**0.93**	0.07	**0.83**	0.21
multi		0.994	**0.01**	0.91	0.09	0.80	0.23
single	10^3	**0.998**	**0.01**	**0.93**	**0.06**	**0.83**	**0.15**
multi		0.997	**0.01**	0.92	0.07	0.77	0.20
single	10^4	0.996	**0.01**	0.89	0.10	0.57	0.54
multi		0.996	**0.01**	0.88	0.10	0.56	0.57

between consecutive poses. For people at small scale and slow motion velocity this is often the case.

Reduction of Frame Rate Finally, in contrast to the scaling approach, we also evaluated the impact of reduced frame rates on the performance of the keypoint tracker. Although the spatial distance between two consecutive poses

Table 4.3: Prediction based on last difference. This table shows the results following a naive approach that takes the last observed difference between timestep $t-1$ and t as a prediction for the next pose at timestep $t+1$. For the reduced frame rate scenarios, the naive approach is inferior to both single- and multi-keypoint approach if α is chosen between 10^1 and 10^3.

Framerate [fps]		**30**		**6**		**3**	
	α	**OKS**	**αMSE**	**OKS**	**αMSE**	**OKS**	**αMSE**
naive	10^0	0.984	0.01	0.85	0.14	0.73	0.38
naive	10^1	0.983	0.02	0.86	0.14	0.74	0.38
naive	10^2	0.988	0.01	0.86	0.14	0.73	0.38

increases with reduced frame rate, the results do not improve. This is most probably mainly due to the loss of information. While scaling increases the overall size of poses and distance between these, it keeps the information of the movement itself. This makes it easier to anticipate the pose for the next timestep. By reducing the frame rate more complex movements between consecutive time steps are conceivable, which makes it more difficult to produce the correct prediction.

Naive Prediction Comparison Since the comparison of the learning-based approaches for single- and multi-keypoint setup showed that the former achieves as good results as the latter, we were also interested whether an even simpler approach could be sufficient for a crowd setup. Therefore, a naive approach was examined that takes its prediction for the next timestep solely based on the difference between the last two time steps. Table 4.3 reports the obtained results for this naive approach and for values of α up to 100. The first thing that stands out are the consistent results over all scales α. This is logical, since we just take the existing pose difference between the last two timesteps and add it on the current pose to obtain a prediction. Smaller deviations are due to the evaluation process which takes random snippets from each sequence in the test set to evaluate the performance. Since the approach is of linear nature, it behaves independent from the scale identical. Concerning the reduction of the frame rate, the same observations could be made as in the single- and multi-keypoint

experiments. This naive approach performs as both examined NPP-approaches at $\alpha = 1$, but starts to get outperformed for bigger values of α.

5 Conclusion

This work examines the suitability of a state-of-the-art human pose tracking method proposed by Okada et al. [6] for application within video surveillance scenarios. We re-implemented the initial algorithm proposed in the paper and applied it to synthetically generated crowd data for training and evaluation purposes. In addition to that, we extended and adapted the method by investigating its performance on whole-body predictions and expanded the inter-timestep distances in two ways. With the single-keypoint approach performing almost all the time best, the assumption came up that simpler predictions perform better in settings with persons at small scale. We therefore finally compared the single-keypoint approach to a naive prediction approach solely based on the last measured offset.
In future work, we will examine the impact further ways to create such keypoint bundles, e.g. on tracking-by-detection using bounding boxes or a combined way to smooth out the resulting tracklets. We furthermore will go from synthetic data to real world data, which to this point has only been evaluated on a qualitative way which was not part of this report.

References

[1] COCO Consortium. *COCO - Keypoint Evaluation*. Accessed: 2020-11-01. URL: `http://cocodataset.org/#keypoints-eval`.

[2] Matteo Fabbri et al. “Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World”. In: *European Conference on Computer Vision (ECCV)*. 2018.

[3] Thomas Golda. “Image-based Anomaly Detection within Crowds”. In: *Proceedings of the 2018 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer, M. Taphanel*. Vol. 40. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, Karlsruhe, 2019, pp. 11–24. ISBN: 978-3-7315-0936-3.

[4] Thomas Golda et al. “Image domain adaption of simulated data for human pose estimation”. In: *Artificial Intelligence and Machine Learning in Defense Applications II*. Ed. by Judith Dijk. Vol. 11543. International Society for Optics and Photonics. SPIE, 2020, pp. 112–127. DOI: 10.1117/12.2573888. URL: https://doi.org/10.1117/12.2573888.

[5] J. Li et al. “CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10855–10864. DOI: 10.1109/CVPR.2019.01112.

[6] Masashi Okada, Shinji Takenaka, and Tadahiro Taniguchi. *Multi-person Pose Tracking using Sequential Monte Carlo with Probabilistic Neural Pose Predictor*. 2020. arXiv: 1909.07031 [cs.CV].

DOE-based Multi-spot Confocal Interference Microscope

Zheng Li

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

Abstract

Diffractive lens arrays are used to enhance the resolution of a low-numerical-aperture objective to perform high-resolution large-area measurements. However, the axial resolution of such setups is still fundamentally limited by the objective[4]. This work introduces the concept of utilizing the reflected conjugate wave of the diffractive optical elements (DOEs) for interference measurements. A traceable step height target is measured and the experiment result shows that the proposed setup can improve the accuracy and reduce the measurement uncertainty in the axial direction.

1 Introduction

Nowadays, with the development of nanotechnologies, there are more and more needs for precise measurement of the small surface structures over a large area, such as semiconductor wafers and meta-surfaces. Confocal microscopy has always been one of the most commonly used methods for surface measurements. It can provide high resolution while being contactless and easy to use. However, microscope objectives have the trade-off between resolution and field of view

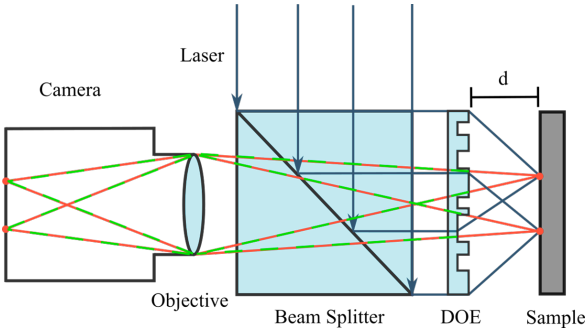


Figure 2.1: Setup of the DOE confocal interference microscope.

(FOV). High-resolution objectives with high numerical apertures (NAs) also have high magnifications, which lead to limited FOVs.

In the previous researches, diffractive lens arrays have been proposed to increase the FOVs of the confocal microscope objectives while maintaining high lateral resolution [3, 6, 5]. Unlike the traditional micro-lens arrays, they no longer have the restrictions between the pitches and the NAs of the micro lenses. They can produce highly focused spots in a dense grid over a large area. Combined with low-NA objectives, large-area measurements with high lateral resolution become possible. However, such setups also have some drawbacks. The axial resolution in this case is still limited by the low-NA objectives.

To overcome such limitation and reduce the axial measurement uncertainties, the idea of interference is utilized. As one of the most precise distance measurement techniques, interference can improve the axial resolution of the previous DOE based confocal microscopes. The concept for the proposed setup will be explained in the next sections. Experiments of a traceable step height target are also carried out to demonstrate such improvements.

2 Confocal Interference Microscope

Figure 2.1 shows the concept of the DOE based confocal interference setup. In the setup, a high-coherence light source, i.e. a collimated laser, is reflected by a beam splitter and illuminates the DOE. Under the plane-wave illumination, the DOE will produce multiple tiny spots with high NAs on the sample surface. These spots will be reflected by the sample as the probe beam, which is shown as red lines in Figure 2.1. At the same time, the DOE will reflect a certain amount of light back as the conjugate wave of the focused spots. The conjugate wave is represented by the green lines in Figure 2.1, which becomes the reference beam. The reflectance of the DOE is determined by the material itself, which is around 4% for fused silica at an incident angle of $90°$ and can be further controlled by a coating.

An objective collects the probe beam and the reference beam and forms an image on the camera sensor. When a surface is placed on the focal plane, the two beams produce two sets of spots in the image, which is shown in Figure 2.2a. Note that the yellow spot is from the direct reflection of the plane surface, which should have the same phase as the probe beam and will not overlap with spots for a sample placed with an angle. By mechanical alignment of the positions of the beam splitter, the objective and the DOE in the setup, the reference and probe spots shown in Figure 2.2b can superimpose with each other perfectly, which means that the reference and probe beams overlap with each other perfectly. Thus the reference and probe beams will interfere with each other. In this case, if the distance between the DOE and the sample is defined as d, then the phase shift for the interference between the probe beam and the reference beam is $4\pi d/\lambda$.

With such a setup, when the surface is scanned axially, interference will occur between the reference and probe beams. Interference fringes can be observed on a confocal peak like modulation signals on a carrier wave. Experiments are shown in the next section for demonstrating such phenomena. It will also be applied in measuring a step height target to reduce the axial measurement uncertainty.

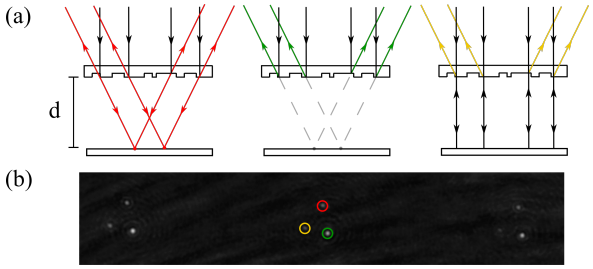


Figure 2.2: Reflection from the DOE and the sample surface. (a) Focusing wave, reflected conjugate wave and reflected wave by the plane surface. (b) Spots from different waves when a piece of glass is placed underneath the setup.

3 Experiment results for surface measurement

For proof of the concept, a testing setup is built as Figure 3.1 shows. A $50\,\mathrm{mW}$ volume-holographic-grating single-frequency $785\,\mathrm{nm}$ fiber-coupled diode laser from Thorlabs is collimated by an objective through a beam splitter. The collimated plane wave illuminates the DOEs, which focuses the illumination into a spot array. The probe beam reflected by the surface and the reference beam reflected by the DOE is collected by the objective and they are imaged onto the camera sensor by the tube lens.

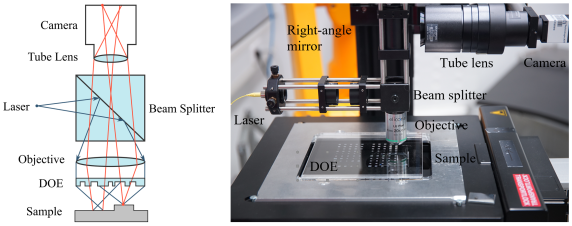


Figure 3.1: Experiment setup of the DOE-based confocal interference microscope.

A traceable step height standard target (VLSI SHS-9400QC) is placed underneath the DOE as the sample to test the measurement uncertainties. The target is shown in Figure 3.2a. There is a convex bar precisely produced by lithography with a width of $100\,\mu\mathrm{m}$ and a calibrated height of $925.5 \pm 5.4\ \mathrm{nm}$. Height measurement of the bar follows the procedures of VLSI [7, 2]. Three lines are measured along the cross section of the bar, which are represented by A, B and C in Figure 3.2b. The length of each line is a third of the width of the bar.

Three adjacent spots in the 11×11 spot array with a pitch of $100\,\mu\mathrm{m}$ produced by the DOE are scanned axially to measure the height of the bar. The DOE has a working distance d of $1\,\mathrm{mm}$. A typical confocal interference signal of a single spot is shown in Figure 3.3a. Interference fringes with oscillation can be observed in the figure. Please note that the fringes disappear on the edge because the axial sampling is set to a larger value to save the measurement time. The highest peak of the signal is fitted by a polynomial and the height is calculated. In this way, each line is sampled with a step of $3\,\mu\mathrm{m}$ in the x direction in Figure 3.2. Afterwards, all the lines are fitted by the following equation [1]

$$z = \begin{cases} ax + b_1, & x \in (A, B), \\ ax + b_2, & \text{otherwise}, \end{cases} \tag{3.1}$$

where z is the measured height, x is the lateral position and a, b_1 and b_2 are the line fitting parameters. The height of the bar h is thus derived as

$$h = \frac{b_2 - b_1}{\sqrt{a^2 + 1}}. \tag{3.2}$$

The measurement is repeated 10 times in the y direction along the bar with a step of $10\,\mu\mathrm{m}$. The bar height at each y position are calculated. The mean value

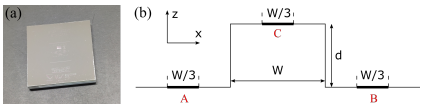


Figure 3.2: VLSI SHS-9400QC step height standard calibration target. (a) Picture of the target. (b) Cross section of the measurement area.

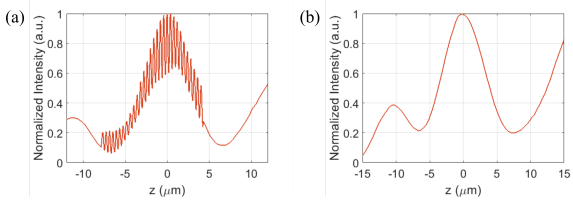


Figure 3.3: Axial measurement signal with a $20\times$ 0.5 NA objective. (a) Confocal interference axial response of a single-frequency laser. (b) Confocal axial response of a laser with a bandwidth of $2\,\mathrm{nm}$.

of the measured height is $904.7\,\mathrm{nm}$ with a standard deviation of $9.7\,\mathrm{nm}$. For comparison, another $785\,\mathrm{nm}$ diode laser from Thorlabs with a larger bandwidth of $2\,\mathrm{nm}$, which has a much shorter coherence length of around $0.3\,\mathrm{mm}$, is used for pure confocal measurement. A typical signal can be seen in Figure 3.3b. Note that the asymmetrical peak shape is due to the aberration of the objective when imaging through the DOE glass, which has a thickness of $6\,\mathrm{mm}$. The confocal measurement has a mean value of $965.5\,\mathrm{nm}$ with a standard deviation of $40.3\,\mathrm{nm}$. It is obvious that the interference measurement significantly reduces the measurement uncertainty.

The measurement error is mainly due to the noise caused by the stray light. The stray light comes from different aspects, such as reflection from the unmatched coating of the objective and other diffraction orders due to the low diffraction efficiency of the binary phase mask. The measurement uncertainty can hopefully be further reduced by changing the objective coating or a multi-level phase mask which has a much higher diffraction efficiency.

4 Conclusion

Microscope objectives with high resolution usually have small FOVs. Diffractive lens arrays can keep both the high lateral resolution and the large FOVs of low-NA objectives in confocal microscopy. However, the axial resolution of such setups are still limited by the low-NA objectives for surface measurement.

This work uses the originally wasted reflected conjugate wave of the DOEs to make interference measurements to increase the axial resolution of previous setups. Experiments are carried out with a calibrated step height target which has a certified height of $925.5 \pm 5.4\,\mathrm{nm}$. The interference measurement show a result of $904.7\,\mathrm{nm}$ with a standard deviation of $9.7\,\mathrm{nm}$. Compared to the solely confocal measurement, which has a result of $965.5\,\mathrm{nm}$ with a standard deviation of $40.3\,\mathrm{nm}$, the measurement uncertainty is clearly reduced and the accuracy is increased.

In the future, a new DOE with higher NAs will be produced to further test the resolution limit of the diffractive lens arrays. New experiment setup will be built and the application will be extended to other fields, for example, fluorescence microscopy.

References

[1] AB Forbes, PM Harris, and Richard K Leach. *The comparison of algorithm for the assessment of Type A1 surface texture reference artefacts.* Tech. rep. CMSC 33/03. National Physical Laboratory, 2018.

[2] Peter de Groot and Danette Fitzgerald. "Measurement, certification and use of step-height calibration specimens in optical metrology". In: *Optical Measurement Systems for Industrial Inspection X*. Ed. by Peter Lehmann, Wolfgang Osten, and Armando Albertazzi Gonçalves Jr. Vol. 10329. International Society for Optics and Photonics. SPIE, 2017, pp. 328–336. DOI: 10.1117/12.2269800. URL: https://doi.org/10.1117/12.2269800.

[3] B. Hulsken, D. Vossen, and S. Stallinga. "High NA diffractive array illuminators and application in a multi-spot scanning microscope". In: *Journal of the European Optical Society-Rapid publications* 7 (2012).

[4] Zheng Li. "Application of diffractive optical elements in confocal microscopy". In: *Proceedings of the 2018 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Relativistic groups and analyticity*. Ed. by Miro Taphanel and Jürgen Beyerer. KIT Scientific Publishing, Karlsruhe, 2019.

[5] Zheng Li et al. “Application of DOE in confocal microscopy for surface measurement”. In: *IMEKO Joint TC1 - TC2 International Symposium on Photonics and Education in Measurement Science*. Ed. by Maik Rosenberger, Paul-Gerald Dittrich, and Bernhard Zagar. Vol. 11144. International Society for Optics and Photonics. SPIE, 2019, pp. 254–261.

[6] Tim Stenau and Karl-Heinz Brenner. “Diffractive Lenses with Overlapping Aperture A New Tool in Scanning Microscopy”. In: *Imaging Systems and Applications*. Optical Society of America. 2016, IT1F–1.

[7] VLSI Standards. *Application Note: Step Height Standards for use with KLA-Tencor Instruments*. Rev.AB. 2010.

A Proposal on Discovering Causal Structures in Technical Systems by Means of Interventions

Josephine Rehak

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

Abstract

Causal Discovery has become an area of high interest for researchers. It has lead to great advances in medicine, in the social sciences and in genetics. But up til now, it is hardly used to identify causal relations in technical systems. This paper presents the basic building blocks for in-depth research. This paper reviews established causal discovery methods and causal models. In contrast to existing surveys of this domain, we focus on the causal discovery methods using interventions. Based thereon, we propose the idea of a promising interventional discovery approach for technical systems. It takes advantage of not only direct, but also indirect causal relationships, which might improve the learning process of causal structures.

1 Introduction

In 360 B.C., Plato already knew that everything in the universe had a cause and is thus an effect of that cause. Nowadays, causality has emerged as the most fundamental common theme in the sciences. Scientists from domains as oceanography, genetics, philosophy or psychology research on discovering

causal relationships. Researching causality has strong prospects for the future, because it can improve human-level artificial intelligence, medicine, precision agriculture or production-optimized factories. In general, learning about causal relations helps to understand technical as well as social systems. By using it, one can better predict possible behavior, alternative outcomes and create policies. Naturally, each domain has its own characteristics. In the social sciences, experiments are rather expensive and are limited by social ethics. Fortunately, we have broader opportunities in the domain of technical systems. They offer the advantage of systematic investigations, while no human lives are at risk. But they also come with new hurdles: Some interventions may not be possible, because a change at the place of investigation is not feasible or could damage the system. We need an approach, which respects these boundaries. The method should efficiently identify causal relationships using a minimal amount of data. In the best case, we can make use of existing algorithms to achieve high knowledge gain with minimal costs. To identify such approaches, we shed light on established methods and causal models. By presenting the current state of research, we join a series of publications such as [7, 8, 13].

2 The Nature of Causal Relations

Causal relations tend to be complex. The most common definition of causality is, that given two events A and B, A causes B, when B relies on A for its value. This causal relation of A and B implies several properties. One is the time-dependency: the event B must not happen before A. B might occur after event A, but the delay time, also called dead time, between A and B must be adhered to. Further on, a causal relation always implies a correlation, but a correlation does not imply a causation. In contrast to correlation, the causal relation guarantees a joint appearance of events A and B. Also the deletion of the cause A, always implies the deletion of B, assumed no other event causes B. Besides the discussed relation $A \rightarrow B$, in the deterministic world there exist four other possible causal relations between the two variables A and B. The easiest one is $A \leftarrow B$, where A depends on B for its value. It is as $A \rightarrow B$ a simple directed relation and we assume them to be the most common relations in nature.

Another possible relation is causal independence denoted as $A \perp B$. In this case, does neither A depend on B, nor B depend on A. It might occur, that the two variables correlate, what is then called a spurious correlation. The trickiest causal relations between two variables are bi-directed causal relations and confounding relations. In a bi-directed relation A and B causally depend on each other. In a confounding relation, A and B are caused by an unmonitored third variable and are indirectly causally related. The whole domain of confounder analysis is solely devoted to finding such relations. Both cases are difficult to detect as they are hard to keep apart from each other and can easily misidentifies as $A \rightarrow B$ or $A \leftarrow B$. Confounding and bi-directed relations have been be denoted as $A \leftrightarrow B$ in the literature. Next to this, there also exists the notation of $A - B$ for bi-directed relations. We will stick to the later in this report.
These were all the possible relations between two variables, but a causal graph with more variables can be much more complex. A variable might actually depend on multiple variables for its value. Because of this, the number of possible models grows exponentially with the number of variables. To reduce this amount of possible models is the main goal of causal discovery methods. In the best case, the true underlying causal model can be uncovered.

3 An Overview of Causal Models

A causal model describes the causal dependencies in a system or population. Causal models can represent only a small part of reality. As every cause has its own cause, the causal network of reality is too gigantic to be fully captured. Hence, a causal model always implies a trade off between complexity and completeness. In literature, the events on the boundary of a model are called exogenous variables. Their value stems from outside of the system. The cause of an event inside the model is called an endogenous event. Since causal models replicate reality, in reverse reality can falsify the models. Different types of causal models have been developed and their number is rising. For these reasons, we limit ourselves to the most popular causal models.

3.1 Structural Causal Models

Structural Causal models (SCMs) were introduced by [24]. They consist of a set of exogenous variables V an a set of endogenous variables W and a set of structural equations F. SCMs belong to the family of Structural Equation Models (SEMs), also called Functional Causal Models (FCM). These kind of models use equations to represent the graph structure. The definition of SEM and SCM is ambiguous. Sometimes SCMs are referred as SEMs. By our definition, SEMs use linear equations, while SCMs use functions. Consequently, SCMs are more powerful than SEMs. In both models, each equation contains an independent noise variable U, which contain the patterns that cannot be causally explained. The probabilistic extension of SCMs are Bayesian Networks.

3.2 Bayesian Networks

Bayesian Networks (BNs), or Belief Networks, belong to the family of Probabilistic Graphical Models (PGMs). They combine graph structures with (conditional) probability distributions. The graph G of a BN consists of a set of random variables X and a set of directed edges E connecting the variables. Per definition, the graph is a directed acyclic graph (DAG), as the edges must not form cycles. Each variable has a probability distribution assigned. According to the Law of Total Probability, all probabilities in a probability distribution sum to one [26]. Popular enhancements of BNs are Dynamic Bayesian Networks (DBNs) [5], Object-Oriented Bayesian Networks (OOBNs) [15].

3.3 Markov Random Fields

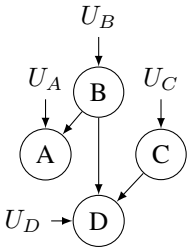
Another way or presenting causal relations are Markov Random Fields [14]. As a PGM, MRFs consist of a set of variables V and a set of probability distributions P. Here, the edge weights do not sum to one. To receive normalized probabilities, we have use the normalization constant Z_ϕ. In opposite to BNs, MRFs contain bidirected edges, which may form cycles. But unidirected relations, as in BNs, are no allowed. Further developments of MRFs are Gaussian Markov Random Fields [29] or Hidden Markov Random Fields [18].

3.4 Acyclic Directed Mixed Graphs

Acyclic Directed Mixed Graphs (ADMGs) were introduced by Judea Pearl [12]. In their essence, ADMGs consist of multiple Bayesian Networks and thus allow bidirected edges. Their only constraint is that unidirected edges are not allowed to form cycles. The bidirected edges stand for a latent common cause, which is not included in the variable set. Successors of the ADMGs are alternative Acyclic Directed Mixed Graphs (aADMGs) [27].

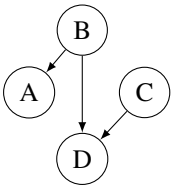
4 Methods for Causal Discovery

The purpose of causal discovery is to recreate the underlying true causal model $G*$ from the set of possible models. This can be done by an algorithm or user posing queries to the data. In observational causal discovery, the query concerns the causal relationship between two variables. A causal discovery method then tries to answer the query using prerecorded data. In interventional causal discovery, the query often concerns the question how the relation between two variables changes, when we intervene on one of them. In this case, the discovery method needs access to a live system which reacts to its interventions. To perform causal discovery, one expects that the Causal Markov assumption holds. This asserts, that the data is generated by an underlying model $G*$ and not by chance. The use of experiments with interventions is the oldest causal discovery approach. Back in 1982 Paul Holland stated: "No Causation without manipulation". He identified interventions as the only method to discover causal relations [10]. Contradicting to this, one could observe a trend to observational causal discovery in the nineties. This trend was crucial for domains where experiments come at high cost and effort. In recent years, there has been a trend towards methods handling a mixture of observational data and data from interventions [22].



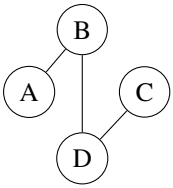
$$
\begin{aligned}
V &= \{A, D\} \quad \textit{endogenous Variables} \\
W &= \{B, C\} \quad \textit{exogenous Variables} \\
F &= \{A = f_A(B, U_A), \quad \textit{Structure Equations} \\
&\qquad B = f_B(U_B), \\
&\qquad C = f_C(U_c), \\
&\qquad D = f_D(C, B, U_D)\}
\end{aligned}
$$

(a) Structural Causal Model



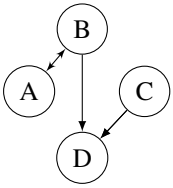
$$
\begin{aligned}
&V = \{A, B, C, D\} \quad \textit{Variables} \\
&E = \{(B \to A), (B \to D), (C \to D)\} \quad \textit{Edges} \\
&P(A, B, C, D) = P(B)P(C)P(A|B)P(D|B, C) \ \textit{Fact.}
\end{aligned}
$$

(b) Bayesian Network



$$
\begin{aligned}
&V = \{A, B, C, D\} \quad \textit{Variables} \\
&E = \{(B - A), (B - D), (C - D)\} \quad \textit{Edges} \\
&P(A, B, C, D) = \tfrac{1}{Z_\phi}\phi(A, B)\phi(B, D)\phi(C, D) \quad \textit{Fact.} \\
&Z_\phi = \textstyle\sum_{A,B,C,D} \phi(A, B)\phi(B, D)\phi(C, D) \quad \textit{Norm.}
\end{aligned}
$$

(c) Markov Random Field



$$
\begin{aligned}
&V = \{A, B, C, D\} \quad \textit{Variables} \\
&E = \{(A \leftrightarrow B), (B \to D), (C \to D)\} \quad \textit{Edges} \\
&f(A, B, C, D) = f(A, B)f(D|B, C) \quad \textit{Factorization}
\end{aligned}
$$

(d) Acyclic Directed Mixed Graph

Figure 3.1: Some of the most prominent causal models in literature are Structural Causal Models, Bayesian Neworks, Markov Random Fields and Acyclic Directed Mixed Graphs.

4.1 Causal Discovery Methods using Interventional Data

This family of discovery methods uses interventions to estimate the direction between two variables. In fact, the term intervention is a cover for many other definitions. It most often stands for perfect interventions, that only affect the desired events and relations. But the term could also be interpreted as imperfect intervention [31], stochastic intervention [16], soft intervention [17], unreliable intervention [6] or uncertain intervention [6]. When we use the term intervention in this report, we refer to perfect interventions.
The most prominent model for interventional deduction is the Potential Outcomes Framework [28]. It was later found to be the main ingredient of Randomized Controlled Trials and A/B tests, the two most popular methods at the time. The PO Framework comes with the do-calculus developed by Judea Pearl [25, 23]. It is a way of denoting, which variables are intervened on. Usually, this method assumes perfect interventions.
To investigate the causal relation between two variables, two randomized sample groups are treated with the potential cause t_1 and an alternative t_2. Then we compare the outcome of both groups by calculating the causal effect $Y_{t_1} - Y_{t_2}$. Imagine an experiment where we want to find out the effect of a drug T. We do this by intervening in the place of the potential remedy $do(T)$. Thereby, T can assume the values 'give medicine' $do(T = t_1)$ or 'give no medicine' $do(T = t_2)$. Then we measure the effects Y_{t_1} and Y_{t_2} in the two independent randomized test groups. For example, the causal effect could be the difference in recovered participants. If the same count of people recover in each test group, the medicine had no effect on the recovery.

4.2 Causal Discovery Methods using Observational Data

This group of methods uses observations of the system behavior. This avoids the effort and cost that come with interventions.
Several groups of methods can learn the causal structure in this way: Common are constraint-based, score-based and other functional methods. Each one makes use of certain statistical patterns in the data. The most important patterns are from conditional independencies [22, 8]. The group of score-based discovery methods

calculates a model fit score from the data. In the next step, it optimizes the score by searching in the space of possible models. Which model the presented method finds is shown in Table 4.1. An example is the Iterative Conditional Mode algorithm [2], which maximizes the fit likelihood of a generated model same as Greedy Equivalence Search (GES). Whereby, GES searches directly in the space of Markov equivalence classes [3], a set of models which express the same conditional independencies.
The constraint-based methods use first conditional independence tests and an edge orientation phase to learn the causal model. Popular are the Inductive Causation (IC) algorithm [33, 24] and the Fast Causal Inference (FCI) algorithm [30]. The disadvantage of such methods result from the conditional independence tests. A conditional independence is not always certain and the tests can require a large amount of data to be faithful.
Besides these groups exists a collection of methods exploiting other properties of causal relations. For example, the Additive Noise Model (ANM) makes use of the independence condition as it only holds true in the causal direction [11]. Opposed to the other method groups, that a false learned relation will not effect all other causal relations of the model [8].
In general, causal discovery based on observations alone can rarely discover the whole true causal graph (ancestral graph). Most often, they are limited to the level of the Markov equivalence. One way to overcome this is by means of interventions [32].

4.3 Causal Discovery Methods using Observational & Interventional Data

In recent years, methods have become popular, which use data created by interventions and observations. Some methods use them in one joint pool and directly construct their causal model from it. While other methods use each kind of data separately to reconstruct the causal graph.
An example pooling method is Joint Causal Inference (JCI) [22]. It is actually a group of methods, as it joins the data from different contexts, preprocesses it and then allows any other common causal discovery method as IC or FCI to construct the causal graph.

Table 4.1: Lists of Observational and Mixed Algorithms that can Discover the Respective Causal Models

Causal Model	Causal Discovery Methods
DAG	ANM [11]
	IC [33, 24]
SCM	JCI [22]
BN	GES [3]
	JCI [22]
MRF	ICM [2]
ADMG	JCI [22]

Other popular pooling methods include [31, 6, 4, 32]. Splitting methods may take diverse forms. The easiest way is to first use observational methods til the Markov equivalence level and then continue by using interventions.

5 Causal Discovery in Technical Systems

In the previous sections, we provided a compact overview over existing methods in causal discovery. Here, we want to give an insight in how we plan on putting them to use in a technical system.
A technical system may take various forms. For example it can be of electrical, water-driven or material-driven nature. Per definition, each artificial system in which matter, power and information interact is a technical system. We assume such a system to be representable in a causal model. From the models described before, ADMGS are the models that capture most domain knowledge.
SCMs and BNs cannot capture bi-directional edges. While MRFs cannot represent the more frequently occurring unidirectional edges [12]. Yet, ADMGs are the least used of the presented models. Further, ADMGs require cost-expensive calculations for causal inference as with each undirected edge, the calculations become more complex [12]. Hence, we advise to first invest in BNs in probabilistic scenarios, and into SCMs or DAGs in non-probabilistic scenarios, before taking the step to ADMGs.

Concerning the discovery methods, we plan to exceed the Markov Equivalence level by using interventions. But as experiments based on the PO Framework examine only a small number of variables [28], it is too expensive to analyze an entire network. Hence, we have to either take observational methods into account or come up with a new method.

6 Research Gap and Proposed Approach

Here, we propose a new form of query which considers such indirect causal relations as of A on C. The basic idea is that we can observe the spread of a change caused by an intervention. To see which intervention has caused which variable changes should enable us to draw conclusions about the causal structure. By collecting several such constraints and by using combinatorial analysis, we hope to deduce the full causal graph. As a side outcome, we might be able to receive such new information as the dead time between cause and effect variables.

The main effort will be detecting the propagating change in the system. We have identified two options. For one, we could compare the state under intervention with a normal state. If the variable of the intervened state deviates from the corresponding variable of the normal state, the variable indirectly depends on the intervention.

In the second option, the intervention creates a kind of signal, which propagates through the system. By trying to recover this signal from the other variables, we detect which variables depend on the intervened variable. We assume this methods to be more difficult, as the signal is likely to change its form when traveling through the system.

For both options, we have to assume consistency in the system environment as external influences on the system can lead to false conclusions in the causal order.

If this new form of query works out, we could use active learning methods to optimize the costs of intervention and the knowledge gain . Several methods such methods for PO queries exist[32, 9]. We would need a new active learning

method for our query.
As a step further in the future, we see the removal of various constraints we impose on the system. For example, we assume our system to be non-cyclic, but also cyclic graphs have been studied in literature [21, 20]. Also we assume to have no influences on the system besides our interventions. It is likely that in a real system this will not be the case and we will have a greater problem in finding created changes.
For the beginning, we will make use of existing simulations of technical systems as the Tennessee Eastman Process [1, 19]. Such simulations are easy to calculate and allow rapid progress in the development of new algorithms. They are already mathematically reproduced and thus causally inferred, what also allows our models to be easily validated. This allows an easy entry into the development of causal discovery methods for technical systems before we roll them out on real plants.

7 Conclusion

In this report, we offered a survey on existing methods in structure learning and causal discovery. If it is convenient, such presented methods could be used for causal inference in technical systems, since there are different possibilities and risks than in domains as the social sciences.
A new approaches for the investigation in observational and interventional causal discovery were proposed. We advice for further investigation to consider indirect causal relations and to use combinatorial inference to learn the causal model structure. To discern the effects of interventions from their environment, we proposed to use comparisons with a most similar twin system or to investigate the injection and tracking of a short signal.

References

[1] Andreas Bathelt, N. Lawrence Ricker, and Mohieddine Jelali. “Revision of the Tennessee Eastman Process Model”. In: *IFAC-PapersOnLine*

48.8 (2015). 9th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2015, pp. 309–314. DOI: 10.1016/j.ifacol.2015.08.199.

[2] Julian Besag. "On the statistical analysis of dirty pictures". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 48 (1986), pp. 259–279. DOI: 10.1111/j.2517-6161.1986.tb01412.x.

[3] David Maxwell Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3 (2002), pp. 507–554.

[4] Gregory F Cooper and Changwon Yoo. "Causal discovery from a mixture of experimental and observational data". In: Morgan Kaufmann Publishers Inc., 1999. DOI: 10.5555/2073796.2073810.

[5] Paul Dagum, Adam Galper, and Eric Horvitz. "Dynamic Network Models for Forecasting". In: *Uncertainty in Artificial Intelligence*. Ed. by Didier Dubois et al. Morgan Kaufmann, 1992, pp. 41–48. ISBN: 978-1-4832-8287-9. DOI: 10.1016/B978-1-4832-8287-9.50010-4.

[6] Daniel Eaton and Kevin Murphy. "Exact Bayesian structure learning from uncertain interventions". In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. Ed. by Marina Meila and Xiaotong Shen. Vol. 2. Proceedings of Machine Learning Research. PMLR, 2007, pp. 107–114. URL: http://proceedings.mlr.press/v2/eaton07a.html.

[7] Clark Glymour, Kun Zhang, and Peter Spirtes. "Review of causal discovery methods based on graphical models". In: vol. 10. Frontiers, 2019, p. 524.

[8] Olivier Goudet et al. "Learning functional causal models with generative neural networks". In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 39–80.

[9] Yang-Bo He and Zhi Geng. "Active Learning of Causal Networks with Intervention Experiments and Optimal Designs". In: *Journal of Machine Learning Research* 9 (2008), pp. 2523–2547. URL: http://jmlr.org/papers/v9/he08a.html.

[10] Paul W Holland. "Statistics and causal inference". In: vol. 81. Taylor & Francis, 1986, pp. 945–960.

[11] Patrik Hoyer et al. “Nonlinear causal discovery with additive noise models”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller et al. Vol. 21. Curran Associates, Inc., 2009, pp. 689–696. URL: https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

[12] Pearl Judea. “Causality: models, reasoning, and inference”. In: 521 (2000), p. 8.

[13] Markus Kalisch and Peter Bühlmann. “Causal structure learning and inference: a selective review”. In: *Journal of Economic Methodology* 11 (2014), pp. 81–98. DOI: 10.1080/16843703.2014.11673322.

[14] Ross Kindermann. “Markov random fields and their applications”. In: *American mathematical society* (1980). DOI: 10.1090/conm/001.

[15] Daphne Koller and Avi Pfeffer. “Object-oriented Bayesian networks”. In: *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence* (1997), pp. 302–313.

[16] Kevin Korb et al. “Varieties of causal intervention”. In: ed. by Yeap WK. Zhang C. W. Guesgen H. Vol. 3157. Springer. 2004, pp. 322–331. DOI: 10.1007/978-3-540-28633-2_35.

[17] Christian Kuehnert, Thomas Bernard, and Christian Frey. “Causal structure learning in process engineering using Bayes Nets and soft interventions”. In: *2011 9th IEEE International Conference on Industrial Informatics*. IEEE. 2011. DOI: 10.1109/INDIN.2011.6034839.

[18] Hans Kunsch, Stuart Geman, and Athanasios Kehagias. “Hidden Markov Random Fields”. In: *Annals of Applied Probability* 5.3 (Aug. 1995), pp. 577–602. DOI: 10.1214/aoap/1177004696.

[19] Carla Martin-Villalba, Alfonso Urquia, and Guodong Shao. “Implementations of the Tennessee Eastman Process in Modelica”. In: *IFAC-PapersOnLine* 51 (2018). 9th Vienna International Conference on Mathematical Modelling, pp. 619–624. ISSN: 2405-8963. DOI: 10.1016/j.ifacol.2018.03.105.

[20] Joris M Mooij et al. “On Causal Discovery with Cyclic Additive Noise Models”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011, pp. 639–647. URL: https://proceedings.neurips.cc/paper/2011/file/d61e4bbd6393c9111e6526ea173a7c8b-Paper.pdf.

[21] Joris Mooij and Tom Heskes. “Cyclic causal discovery from continuous equilibrium data”. In: *arXiv:1309.6849* (2013).

[22] Joris Mooij, Sara Magliacane, and Tom Claassen. “Joint causal inference from multiple contexts”. In: *arXiv preprint arXiv:1611.10351* (2016).

[23] Judea Pearl. “Causal diagrams for empirical research”. In: *Biometrika* 82 (1995).

[24] Judea Pearl. *Causality*. Cambridge university press, 2009. DOI: 10.1017/CBO9780511803161.

[25] Judea Pearl. “The do-calculus revisited”. In: *arXiv:1210.4852* (2012).

[26] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[27] Jose Peña. “Alternative Markov and causal properties for acyclic directed mixed graphs”. In: *arXiv preprint arXiv:1511.05835* (2015). DOI: 10.1023/A:1017445827962.

[28] Donald B Rubin. “Causal inference using potential outcomes: Design, modeling, decisions”. In: *Journal of the American Statistical Association* 100 (2005), pp. 322–331. DOI: 10.1198/016214504000001880.

[29] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall, 2005. DOI: 10.1201/9780203492024.

[30] Pater Spirtes et al. “Constructing Bayesian network models of gene expression networks from microarray data”. In: (2000).

[31] Jin Tian and Judea Pearl. “Causal Discovery from Changes”. In: (2001), pp. 512–521. DOI: 10.5555/2074022.2074085.

[32] Simon Tong and Daphne Koller. “Active learning for structure in Bayesian networks”. In: *International joint conference on artificial intelligence*. Vol. 17. 2001, pp. 863–869.

[33] Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

A Step Towards Explainable Person Re-identification Rankings

Andreas Specker

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

Abstract

More and more video and image data is available to security authorities that can help solve crimes. Since manual analysis is time-consuming, algorithms are needed that support e.g. re-identification of persons. However, person re-identification approaches solely output image rank lists but do not provide an explanation for the results.

In this work, two concepts are proposed to explain person re-identification rankings and a qualitative evaluation is conducted. Both approaches are based on a multi-task convolutional neural network which outputs feature vectors for person re-identification and simultaneously recognizes a person's semantic attributes. Analyses of the learned weights and the outputs of the attribute classifier are used to generate the explanations.

The results of the conducted experiments indicate that both approaches are suitable to improve the comprehensibility of person re-identification rankings.

1 Introduction

The increased use of surveillance cameras to ensure security in public spaces leads to huge amounts of video data available to law enforcement agencies. On the one hand, this allows the search for specific persons of interest, but on the other hand, it raises the problem of efficient and fast evaluation of the data.

The research field of person re-identification (re-id) addresses this problem by developing approaches that enable automatic searches for persons in a huge image or video database, usually referred to as the gallery. The starting point for a search is typically a so-called query image that shows the target person.

Recent works [4, 11, 20] train a convolutional neural network (CNN) to embed person images into a feature space. This feature space has the characteristics that generated features from images showing the same person are closer together than features from images of different people. Such features, also called embeddings, are represented by vectors with a certain number of elements N. The calculation of the distances between the query embedding vector and all embeddings of images included in the gallery makes it possible to create a ranking of the gallery images sorted by their similarity to the query image.

Task-specific problems that make it difficult to train a CNN for re-id include low image resolution, occlusions, and misaligned person detections. Moreover, large differences between scientific datasets, typically used for developing and training, and real-world data lead to problems. Scientific datasets are only a small excerpt of reality restricted with respect to the variety of persons' visual appearances. They are usually recorded within a short period and at a specific location and scene setup. As a result, many important characteristics of a person's visual appearance, such as different types of clothing in summer and winter or varying lighting conditions, are not included. Hence, the learned feature space is biased and thus imperfect, especially since it is a matter of finding unseen persons in the application who may have unfamiliar characteristics.

Therefore the resulting rankings of the person pictures naturally contain false positive results in the first ranks as well. In this case, the difficulty is that these errors are not necessarily understandable. The reasons for images being ranked at their positions remain unclear.

Since the search is based ond abstract deep feature vectors, it is not possible to intuitively interpret the embeddings. Furthermore, the sole indication of the distance does not solve the problem since such values are not intuitively interpretable either.

To provide meaningful explanations along with ranking lists, re-id embeddings are first thoroughly analyzed in this work. Subsequently, two concrete approaches for explaining rankings are proposed and evaluated. The general concept behind both approaches is to leverage person attributes, such as gender or clothing colors, to add semantics to the re-id model. Thus, the meaning of the feature vectors can be understood to a certain extent via analyzing the relationship between elements of the embedding vector and the attributes.

2 Related work

This technical report is in the realm of two different research fields: person re-identification and explainable artificial intelligence (XAI). As this work can be applied to any person re-id approach, related work regarding re-id will not be discussed further in this section. According to [18], XAI methods can be categorized into three main fields: explaining of inner-workings, counterfactual explanations, and explanation of decisions.

One possible way to explain the **inner-workings** of a CNN is to determine and visualize features that maximize the activation and are thus most relevant [2, 14, 8]. Other recent works focus not on the maximization of activation but instead try to invert neural networks and retrieve explanations based on e.g. parameter gradients [5, 12, 19]. Moreover, some approaches distill the information of deep neural networks into models with better interpretability [17, 10] or aim to characterize hidden features quantitatively [1, 13].

Counterfactual explanations in the context of XAI describe what has to be changed to the feature vector in order to achieve a prediction of the desired class. For example, such explanations can have the form *"If feature value X would be Y, class C would have been predicted"*. Works that investigate counterfactual explanations to achieve more interpretable machine learning models are [21, 6, 9].

The work in this technical report best fits the research direction of **explaining decisions**. To do this, most approaches rely on the visualization of attribution maps, such as gradient or activation maps, or leverage attention modules to generate valid explanations. Commonly used methods are [16, 3, 7]. In contrast to these methods which primarily focus on the explanation of classification results, person re-id is a retrieval task and does not make any hard decision. Instead abstract feature vectors are compared. To bridge this gap, this work adds an attribute classifier in order to be able to make differences between feature vectors from hidden layers more interpretable.

3 Concepts

The main idea behind the proposed concepts is to use a pre-trained re-id network as a black box and to train an attribute classifier upon it. The attribute classifier takes re-id embeddings as input and outputs classification probabilities of the recognized attributes. The parameters of the re-id network remain frozen while only the weights of the newly added fully connected classification layer are trained. By that, the attribute classifier is forced to interpret the abstract feature vectors and to recognize the attributes based on the information contained therein. The architecture is visualized in Figure 3.1.

Of course, this training procedure does not achieve the best results in terms of attribute recognition accuracy, but it allows the interpretation of the meaning of the elements of the re-id feature vector. The learned weights of the fully connected attribute classification layer enable direct conclusions to be drawn between feature components and their meaning concerning semantic attributes. The weights are understood as a measure of the correlation between the embedding and the attributes.

3.1 Use of classifier outputs

The straightforward way to explain ranking results is to compare attribute predictions of query and gallery images. It would be possible to compute the distances between attribute predictions and to output those for each attribute

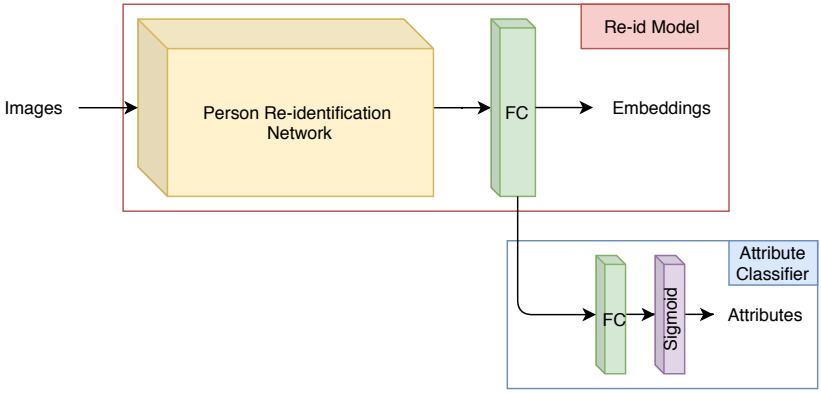


Figure 3.1: Visualization of the multi-task network architecture used in this work. The CNN jointly generates person re-identification embeddings and recognizes semantic attributes.

to explain the matching, but this method suffers from several drawbacks. First, depending on the number of attributes, displaying scores for all attributes could overstrain the system operator because they can not be captured and understood at a glance. Additionally, some attributes might not be visible or relevant to re-identify the person shown in the query image and thus do not need an explanation. Furthermore, absolute errors are hard to interpret without expert knowledge and reference values. As a result, it would be beneficial to have matching scores in percent instead. Building on the identified problems, the following method is proposed to generate meaningful clues on the positions of gallery images. Since the goal is to find occurrences of the person visible in the query image, the first step is to identify the attributes for which the classifier is most certain. Confidently recognized query attributes are determined by computing the distances between the classifier outputs and the attribute decision boundary as a measure of uncertainty. Typically, the decision boundary is 0.5. Afterward, a decision is made based on a threshold t_a. So, with x being the classifier output for attribute a, attribute a is chosen if $|x_a - 0.5| > t_a$ applies. For example, for $t_a = 0.1$ attributes with classification scores below 0.1 or above 0.9 would be selected as suitable candidates to help to explain the ranking results. In the next step, the absolute errors between query and gallery images

are computed for these attributes. As the aim is to provide matching instead of error scores, the error measurements have to be inverted. Last but not least, normalization by the attribute prediction confidence of the query image results in matching scores in percent. The following equation points out the computation formula in detail.

$$s_a = \frac{1 - |x_a^q - x_a^g|}{0.5 + |x_a^q - 0.5|} \tag{3.1}$$

Here, q denotes the query image and $g \in G$ stands for a gallery image from gallery G.

3.2 Attribute-related error

The second concept for explaining the person rank list focuses on the retrieval distance instead of attribute classifications. The goal is to visualize the contribution of each attribute to the distance between q and g. This approach exposes the attributes which contribute most to the distance between the embedding vector. To achieve this, the squared error for each element of the query feature vector f^q and the gallery feature vector f^g is multiplied with the learned attribute classifier weight w_{na}. w_{na} denotes the weight between feature component n and attribute output neuron a. As can be seen in the following Equation 3.2, summing the weighted errors for all feature elements results in an error measurement e_a for attribute a.

$$e_a = \sum_{n=1}^{N} (f_n^q - f_n^g)^2 * w_{na} \tag{3.2}$$

Comparing the errors of the attributes allows the estimation of their contribution to the retrieval distance.

4 Evaluation

This chapter focuses on two main aspects. First, embeddings for person images are analyzed to understand the influence of single feature elements and to examine the correlation with semantic attributes. The more individual vector

elements correlate with single or few attributes, the easier it is to understand and interpret their values and to explain ranking results.

Second, the proposed concepts for explaining the resulting rank list of gallery images are evaluated qualitatively based on some meaningful examples.

4.1 Training and Parameters

All experiments presented in this work are conducted with the well-known Market-1501 [22] dataset as it provides identity labels as well as annotations for 27 attributes. The dataset consists of $32,668$ images of $1,501$ persons, divided into training, query, and test sets. For the experiments, the multi-class attribute *age* is also assumed to be binary, resulting in 30 binary attributes.

For the experiments, the AGW approach [20] is used as the person re-id model. It achieves results comparable to the state-of-the-art with a simple architecture. It is trained using the standard parameters proposed in the original work.

Subsequently, the network parameters of the re-id model remain frozen while a fully-connected classification layer appended to the re-id feature layer is trained. This additional layer consists of 2048×30 weights which equals the number of re-id vector elements times the number of binary attributes in the datasets. Please note that learning a bias is omitted in this layer. Regarding training parameters and procedure, this work orients itself on the findings of [15].

4.2 Embeddings and corresponding attributes

To create Table 4.1, the learned weights between each of the 2048 components of the fully-connected feature layer and the attribute classification layer were examined. For each component, the attribute with the highest weight was determined and summed up with respect to the attributes. For instance, 104 vector components have the greatest weight with attribute *downblue*. The third column refers to the positive ratio of attributes in the training dataset since obviously there is a relationship between positive ratios and the number of top-1 occurrences. The results in the table indicate that the problem of imbalanced or biased data is not only limited to the task of person attribute recognition. Persons

with rarely occurring attributes such as *hat* or *downyellow* are worse represented in feature space and thus the error probability of the resulting ranking increases. This is a particular problem when it comes to ethnically unbalanced training data. Besides, the second factor that is relevant for the number of neurons connected to attributes is the complexity of the attribute. For example, the attribute *bag* occurs in lots of different types, colors, and styles. Thus multiple feature elements are required to represent such a huge variety.

Table 4.1: The number of vector elements that have the greatest weight to the attributes compared to the positive ratios of attributes in the training dataset.

Attribute	#Top1	Positive Ratio
bag	122	24.63
age2	104	75.77
downblue	104	16.38
backpack	103	26.50
downgray	101	16.38
...	...	...
age4	39	1.07
downgreen	33	1.86
downyellow	32	1.33
downpurple	15	0.27
hat	8	2.66

4.3 Attributes and corresponding features

Next, it is examined what the individual elements of the embedding represent. Since there are connections between all feature elements and attributes, single elements likely represent combinations of several attributes rather than single attributes. Figure 4.1 shows rank lists of gallery images for meaningful and representative feature vector elements sorted by their values. The first column contains the three attributes with the highest and lowest weights, respectively. It

is noteworthy that since weights can have negative values, the attributes with low weights are inhibited if the respective vector element has a high value.

Figure 4.1: Rank lists of gallery images for selected vector elements. The images are sorted in descending order by the value of the corresponding embedding element.

The rank lists indicate that single embedding components stand for combinations of attributes. For example, all persons with high values for the element wear yellow shirts and belong to age group 2 (*upyellow, up, age2*) in Figure 4.1(b). Besides, if the color of the upper-body clothing is yellow, it is reasonable to inhibit the prediction probability of the attribute *upblack*. Another interesting finding is that many components not only stand for an attribute combination but are almost able to identify specific persons, like in Figure 4.1(a). Possible explanations are that the dataset consists of fewer identities than elements in the embedding vectors or that the last layer of a re-id network is already very focused on different persons and not concepts similar to attributes. The second explanation is in line with the results of [15]. The authors achieved the best results if the last network layer is not shared between the re-id and attribute recognition tasks in a multitask network. The results showed that there is interference due to different training goals. Features from the last layer of a re-id network are too focused on different identities, even with the same set of semantic attributes.

4.4 Towards understanding rankings

This section provides examples of the proposed concepts to explain rankings based on semantic attributes. First, Figure 4.2 shows ranked gallery images with certain query attributes and corresponding matching scores for images in the top-10 ranks. Second, the error composition for the last sample is discussed based on Figure 4.3.

The first example in Figure 4.2(a) shows a frontal view of the person of interest in good quality. As a result, many attributes are reliably recognized. These attributes cover all types of semantic attributes ranging from global ones like *gender* and *age* over accessories to clothing styles and colors. At first glance, early ranks only contain persons that share a very similar visual appearance. However, there are also cases of error. For example, when looking at rank 8, one can observe that the image shows a woman although a man is visible in the query image. Concerning the proposed matching scores, it gets clear why this image occurs in an early rank. The matching scores for the attributes are high even for the *gender* attribute (! denotes that an attribute is not present. In this case, *gender* means female while *!gender* stands for male). It can be concluded that the CNN was not able to recognize correctly that the image on rank 8 shows a woman in contrast to the query image. This is indeed a difficult case because there are few clues. For instance, the long hair of the person is hardly visible from the frontal view and could also be a part of the background.

In contrast to the first example, the query image of the second example (see Figure 4.2(b)) is blurry and shows heavy occlusions. As a result, the re-id network is unsure about most of the attributes. For example, it is noticeable that irrelevant features such as the concealing bicycle are re-identified instead of the person. Together with the fact that even early ranks only have matching scores significantly below 75%, it indicates that the result is not very reliable. In practical application, the system operator should use another query image with better quality and fewer occlusions in such a case.

Regarding the example in Figure 4.2(c), clothing colors are certainly predicted, but global attributes like *gender* can not be surely determined. As a result, early ranks contain persons with the same combination of short, red upper-body clothing and black trousers with high matching scores. Furthermore, top-10

age2	99.94%	99.74%	99.55%	97.70%	99.29%	99.69%	96.87%	99.02%	97.41%	96.77%
backpack	97.99%	97.44%	93.78%	97.45%	95.37%	97.03%	95.05%	96.52%	92.61%	98.71%
!bag	96.10%	76.44%	99.40%	90.47%	68.91%	88.77%	99.55%	96.24%	97.58%	96.50%
clothes	99.36%	99.20%	99.35%	98.82%	97.98%	98.38%	98.03%	95.57%	93.49%	98.27%
down	96.96%	99.77%	96.07%	95.69%	92.65%	94.80%	96.82%	96.66%	97.80%	81.16%
downgray	98.97%	99.86%	98.12%	100.00%	99.09%	99.61%	99.16%	93.70%	98.55%	90.94%
!gender	99.01%	95.86%	96.52%	97.50%	98.64%	97.34%	92.97%	98.54%	97.23%	79.84%
!hair	97.19%	95.12%	97.03%	96.96%	93.12%	96.65%	98.75%	94.02%	96.36%	90.69%
upwhite	99.56%	99.11%	97.56%	98.45%	97.91%	98.60%	98.52%	97.86%	94.31%	97.54%

(a)

age2	96.87%	95.88%	91.07%	93.52%	85.64%	67.28%	86.51%	96.37%	95.92%	63.54%
upwhite	99.58%	97.75%	74.79%	97.43%	99.16%	89.60%	97.35%	98.24%	98.27%	98.13%

(b)

clothes	95.45%	98.48%	98.76%	91.01%	91.23%	98.37%	95.09%	98.65%	80.96%	99.48%
downblack	99.47%	99.53%	99.17%	99.72%	99.76%	98.96%	98.81%	99.34%	95.99%	99.09%
upred	96.31%	99.11%	99.96%	98.21%	96.84%	89.67%	96.38%	97.64%	98.40%	97.39%

(c)

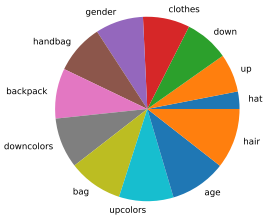
down	98.88%	98.37%	98.12%	99.53%	98.69%	98.29%	96.61%	98.65%	97.36%	99.79%
downgray	99.43%	98.14%	93.00%	98.09%	97.65%	95.31%	91.44%	95.13%	93.61%	99.58%
gender	99.92%	99.69%	99.89%	99.52%	99.46%	99.14%	99.72%	99.51%	99.38%	99.52%
hair	99.86%	99.48%	98.22%	99.36%	99.21%	99.76%	98.76%	97.49%	88.84%	98.20%
!hat	99.18%	97.68%	99.05%	94.70%	94.28%	96.45%	98.78%	96.18%	97.01%	95.97%

(d)

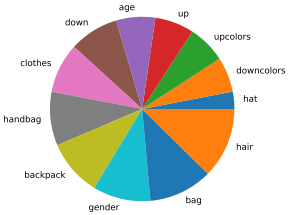
Figure 4.2: Examples for the proposed matching scores based on semantic attributes. ! before an attribute indicates that the attribute is not present in the image.

ranks include both men and women, since gender is not clear from the query image.

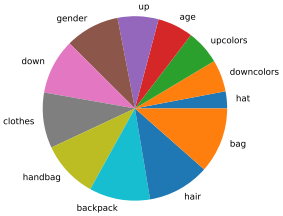
The last example again contains only gallery images with high matching scores for certain query attributes. Except for the last picture, the top-10 images show the person of interest. However, the tenth rank visually differs significantly from the target person. The woman wears a yellow shirt in contrast to the dress the query person is wearing. This example is used to look into detail regarding the error or distance composition as explained in Section 3. Figure 4.3 visualizes error composition for images on rank 3, 9, and 10 as pie charts.



(a) Rank 3



(b) Rank 9



(c) Rank 10

Figure 4.3: Error composition for three different rank images for the rank list shown in 4.2(d)

It attracts attention that the upper-body color greatly contributes to the retrieval distance for rank 3, but less for ranks 9 and 10. The reason for this is that backpacks cover large parts of the upper body clothing when a person is visible from behind. This fact explains the early position of the image on the tenth rank. Matching is not done by the actual color of the upper-body clothing. Instead, the network focuses on the color of the backpack and does not notice the yellow shirt. As a result, the proposed concept allows the understanding of re-id rankings and enable easier identification of weaknesses of the re-id approach used.

5 Conclusion

This work presented concepts to explain and understand rank lists of person re-id system. For this, an attribute classifier is trained with the goal of adding interpretable semantics. Qualitative evaluations show that the proposed concepts work and provide a solid basis for explainable person re-id. It seems to be worth it to conduct further investigations in future work. Interesting research directions include the development of methods for quantitative evaluation as well as generative adversarial networks (GAN). GANs would allow the manipulation of certain aspects of a person's visual appearance and thereby an examination of the effects and influences on ranking results.

References

[1] Chirag Agarwal, Peijie Chen, and Anh Nguyen. "Intriguing generalization and simplicity of adversarially trained neural networks". In: *arXiv preprint arXiv:2006.09373* (2020).

[2] Santiago A Cadena et al. "Diverse feature visualizations reveal invariances in early layers of deep neural networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 217–232.

[3] Aditya Chattopadhay et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 839–847.

[4] T. Chen et al. “Abd-net: Attentive but diverse person re-identification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 8351–8361.

[5] Alexey Dosovitskiy and Thomas Brox. “Inverting visual representations with convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4829–4837.

[6] Yash Goyal et al. “Counterfactual visual explanations”. In: *arXiv preprint arXiv:1904.07451* (2019).

[7] Dong Huk Park et al. “Multimodal explanations: Justifying decisions and pointing to the evidence”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8779–8788.

[8] Qi Li et al. “Improving sample diversity of a pre-trained, class-conditional GAN by changing its class embeddings”. In: *arXiv:1910.04760* (2019).

[9] Shusen Liu et al. “Generative counterfactual introspection for explainable deep learning”. In: *arXiv preprint arXiv:1907.03077* (2019).

[10] 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.

[11] Hao Luo et al. “Bag of tricks and a strong baseline for deep person re-identification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.

[12] Aravindh Mahendran and Andrea Vedaldi. “Understanding deep image representations by inverting them”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5188–5196.

[13] Uday Singh Saini and Evangelos E Papalexakis. “A Peek Into the Hidden Layers of a Convolutional Neural Network Through a Factorization Lens”. In: *arXiv preprint arXiv:1806.02012* (2018).

[14] Shibani Santurkar et al. “Computer Vision with a Single (Robust) Classifier.” In: (2019).

[15] Andreas Specker, Arne Schumann, and Jürgen Beyerer. “A multitask model for person re-identification and attribute recognition using semantic regions”. In: *Artificial Intelligence and Machine Learning in Defense Applications II*. Ed. by Judith Dijk. Vol. 11543. International Society for Optics and Photonics. SPIE, 2020, pp. 98–110. DOI: 10.1117/12.2573981. URL: https://doi.org/10.1117/12.2573981.

[16] Jost Tobias Springenberg et al. “Striving for simplicity: The all convolutional net”. In: *arXiv preprint arXiv:1412.6806* (2014).

[17] Sarah Tan et al. “Distill-and-compare: Auditing black-box models using transparent model distillation”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 303–310.

[18] Giulia Vilone and Luca Longo. *Explainable Artificial Intelligence: a Systematic Review*. 2020. arXiv: 2006.00093 [cs.AI].

[19] Eric Wong and J Zico Kolter. “Neural network inversion beyond gradient descent”. In: ().

[20] Mang Ye et al. “Deep Learning for Person Re-identification: A Survey and Outlook”. In: *arXiv preprint arXiv:2001.04193* (2020).

[21] Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. “Interpreting neural network judgments via minimal, stable, and symbolic corrections”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 4874–4885.

[22] Liang Zheng et al. “Scalable Person Re-Identification: A Benchmark”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

Multi-Object Tracking in Drone Videos

Daniel Stadler

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

Abstract

In this report, three popular methods for multi-pedestrian tracking are extended to a multi-category setting and tested on a large drone-based dataset. A thorough comparison of the algorithms is presented and a common shortcoming is identified. Building on this, a new tracking-by-detection based approach is developed that outperforms the other methods by a large margin. In addition, a state-of-the-art object detection model is adapted for the drone imagery, since no public detections are available for the dataset.

1 Introduction

Multi-object tracking (MOT) in drone videos has several applications ranging from sports analysis to traffic surveillance. Challenges arise not only from complicated scenes with occlusions or fast moving objects but also from different camera altitudes and angles resulting in a large variance of object size and appearance. To solve these problems, most MOT approaches follow the tracking-by-detection paradigm, where the tracking of objects is divided into two subtasks – detection and association. This procedure has the advantage that the vast improvements of deep learning based object detectors from the last years can be directly applied. After detecting objects in each frame of

a video independently, the goal of the subsequent association step is to link detections of the same object to tracks with a unique ID. This is the part, where the main differences of existing MOT frameworks lie. While some simple methods only use the bounding box information of the detections [2, 3, 4], other more sophisticated approaches use separate networks adopted from person re-identification to extract appearance features of the underlying objects [1, 14]. Further ideas that emerged from the analysis of persons are to use human pose information [7] or the interaction of people [10] to assist the association of detections to tracks. Apart from that, the movement of objects often is considered with a motion model (MM) following a simple constant velocity assumption (CVA) between two consecutive frames when the sampling rate is high or by applying a Kalman filter, for example.

In this report, three approaches originally designed for multi-person tracking are extended in order to treat multiple object categories. After a qualitative comparison, a new tracking method is developed upon the identification of a common shortcoming of the existing approaches. Furthermore, a state-of-the-art object detector is trained to boost the performance of the applied tracking-by-detection based methods and to allow a fair comparison. The superior performance of the proposed tracker is shown through experiments on a large drone-based video dataset.

2 State-of-the-Art MOT

In this chapter, three popular MOT frameworks originally developed for person tracking are described and extended to the multi-category context. Whereas the first two follow the predominant tracking-by-detection paradigm, the latter one proposes a new concept to perform the association step implicitly. After a qualitative comparison, a new approach is developed that builds upon the weaknesses of the existing approaches.

2.1 Tracking-by-Detection Approaches

One of the simplest trackers is the IOU tracker [3]. It calculates the Intersection over Union (IoU) of all possible assignments and follows a greedy matching strategy that starts to link the detection and track with the highest IoU. Whereas in this tracker, no visual information is used, the advanced V-IOU tracker [4] leverages a KCF [8] as single object tracker to bypass missing detections. Since the KCF works class-agnostic, the V-IOU tracker can be directly applied in a multi-category setting, linking only detections from the same class together.

As a second tracking-by-detection based method, Deep SORT [14] is chosen, since it is one of the most popular MOT frameworks. As a further development of the SORT algorithm [2], it models the movement of objects with a Kalman filter and uses the motion information by calculating the Mahalanobis distance between predicted Kalman states and new detections. However, this motion metric is not used directly in the association step but only to restrict the area of possible matches for a track, i.e., as a gating function. For associating detections to tracks within their gating area determined by the motion metric, appearance features are extracted using a CNN adopted from person re-identification. The visual features of a track and a detection are compared and if their distance is below a threshold they can be linked. Instead of a greedy matching strategy, the Hungarian algorithm [9] is used for an optimal assignment. Although the separate model for generating appearance features was only trained on person data, it is found that the extracted features are also suitable for comparing objects of other categories like cars or buses. Therefore, Deep SORT also can simply be extended to track multiple categories.

2.2 Tracktor

In contrast to the aforementioned trackers, Tracktor [1] goes beyond the tracking-by-detection procedure. Detections of the previous frame are used as additional region proposals in the second stage of a Faster R-CNN [12] detector and regressed to the new positions in the current frame. Hence, no association step is needed and the tracking is done implicitly. Tracks are stopped if the score of the classification branch falls below a threshold and the detector runs in

parallel to start additional tracks when new objects appear in the scene. Since the regression does not work when large object displacements are present, the method is extended with a CVA as MM and a camera motion compensation (CMC) model that applies the Enhanced Coefficient Maximization technique from [6]. Additionally, a re-identification model checks for new detections whether they belong to earlier interrupted tracks, so that occlusions can be bypassed, leading to the improved version Tracktor++. Although the method is developed for tracking persons, it can be extended to multi-category tracking. For this, the score vector of the classification branch is taken when regressing a track and if its maximum does not correspond to the same category as in the previous frame the track is stopped.

2.3 Own Approach: PAS Tracker

The goal is to build a tracker that uses as much information as possible in the association step and combines the different cues of objects like **p**osition, **a**ppearance and **s**ize in a sophisticated way in order to use all of the information at the same time. In contrast, the previously described algorithms do not use all available information and apply one cue after another: The V-IOU tracker considers appearance information only in the post processing basing the association solely on IoU, whereas Deep SORT takes position information only for gating and relies mainly on appearance features for linking detections. Tracktor++ takes appearance of objects only into account to retrieve lost tracks, but does not use this source of information in the association step.

To overcome the aforementioned limitations, the similarity measure between the position of a detection and a track should fulfill the following three requirements. First, the center of objects shall be directly compared instead of using the IoU, which is not accurate enough for densely packed small objects as often present in the drone context. Second, a similar gating mechanism to inhibit impossible matches as in Deep SORT is desirable. Therefore, the position similarity has to be zero for too large displacements. Third, in order to enable a straightforward combination with other similarity measures, the metric should be normalized between zero and one. Given the position of a detection $\mathbf{p}_D$ and the position of a track $\mathbf{p}_T$ (after MM and CMC) with center coordinates $\mathbf{p} = (x, y)$, the

position similarity s_p is calculated as follows:

$$s_\mathrm{p} = \max(1 - \lambda_\mathrm{p}||(\mathbf{p}_T - \mathbf{p}_D) \oslash \mathbf{z}_D||, 0) \tag{2.1}$$

$\oslash$ denotes the element-wise division, $||$ the Euclidean norm and $\mathbf{z}_D = (w, h)$ the size of the detection, i.e. the width and height of the bounding box. The normalization w.r.t. the object size accounts for varying camera altitudes that lead to differently large displacements in the image. The hyperparameter λ_p tunes the size of the gating area, where the position similarity is not zero. A good choice of this value is related to the displacements of objects between frames, thus to the velocity of the moving objects and the camera frame rate ($24fps$ in the VisDrone MOT dataset [15]); λ_p is empirically set to 0.3.

As a second similarity measure, the size of objects is compared. Similar to the position information, the IoU also reflects size similarity but is not very accurate, since it does not measure position and size similarity independently. To get a maximum similarity score of one, the following formula to calculate the size similarity s_z is used:

$$s_\mathrm{z} = 1 - ||(\mathbf{z}_T - \mathbf{z}_D) \oslash (\mathbf{z}_T + \mathbf{z}_D)|| \tag{2.2}$$

For a visual comparison of detections and tracks, the improvements from the person re-identification community are leveraged using a state-of-the-art model from [11]. With this model, 2048-dimensional feature vectors are extracted for a detection $\boldsymbol{\theta}_D$ or a track $\boldsymbol{\theta}_T$. Then, the appearance similarity s_a is calculated as cosine similarity like in the Deep SORT framework:

$$s_\mathrm{a} = \frac{\boldsymbol{\theta}_T \cdot \boldsymbol{\theta}_D^\mathrm{T}}{||\boldsymbol{\theta}_T|| \cdot ||\boldsymbol{\theta}_D||} \tag{2.3}$$

Since s_a also gets one for maximum similarity, the three aforementioned metrics can be easily combined to a joint similarity measure s in order to use all the information at the same time in the association step:

$$s = s_\mathrm{p} \cdot s_\mathrm{a} \cdot s_\mathrm{z} \tag{2.4}$$

This similarity is calculated for each track-detection pair and an optimal assignment is achieved with the Hungarian algorithm. A CVA is taken as MM, since

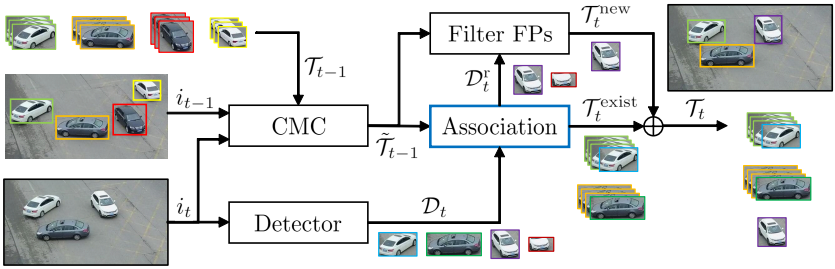


Figure 2.1: Workflow of the proposed PAS tracker [13]. A detector takes as input the current frame i_t and generates a set of detections $\mathcal{D}_t$. Then, a CMC model calculates a transformation using the current frame i_t and the previous frame i_{t-1}. This transformation is applied on the previously found tracks $\mathcal{T}_{t-1}$ yielding the tracks $\tilde{\mathcal{T}}_{t-1}$ that are aligned with the current frame. For each track-detection pair, a similarity measure is calculated and the detections with a high similarity are assigned to the existing tracks $\mathcal{T}_t^{\text{exist}}$. Before starting new tracks $\mathcal{T}_t^{\text{new}}$, the remaining un-assigned detections $\mathcal{D}_t^r$ go into a module that filters false positive detections.

in drone-based imagery usually a large frame rate is available and a Kalman filter yielded no better results in the experiments. To compensate for fast camera movements, the same CMC model as in Tracktor++ is adopted. As a further component, a simple yet effective module to filter false positive detections in crowded scenarios is introduced, since these cause many ID switches. For each new detection that has not been assigned in the association step, its overlap with existing tracks is computed and the detection is removed if the overlap is too high (>0.8) arguing that it is unlikely for objects to appear at positions where already other tracks are present. The complete workflow of the PAS tracker is visualized in Figure 2.1.

3 Experiments

At first, an overview of the dataset on which the presented tracking methods have been evaluated is given. Next, the applied object detector and the adaptations made to cope with the drone-based imagery are described. Finally, the results of the conducted experiments are presented.

Figure 3.1: Example images of the VisDrone MOT dataset [15].

3.1 Dataset

To analyze the performance of state-of-the-art MOT methods in the context of drone imagery, a suitable dataset is needed. For this purpose, the VisDrone MOT dataset [15] is chosen, since it is the largest drone-based dataset for MOT that is publicly available. The dataset consists of 96 videos comprising about 40,000 frames with resolutions up to 3840 $\times$ 2160 pixels and is divided into 4 splits – *train* (56), *val* (7), *test-dev* (17) and *test-challenge* (16). Note that the annotations of the test-challenge split are hidden and used for a yearly challenge hosted by the VisDrone team. Therefore, the test-dev split is used for evaluation. The five categories *pedestrian*, *car*, *van*, *truck* and *bus* are evaluated as it is done in the challenge. Figure 3.1 shows some example images of VisDrone MOT. The dataset is very challenging due to a high variance in camera altitude and viewing angle leading to diverse object appearances and sparse object distributions. Furthermore, both day and night scenes exist.

3.2 Object Detector

Since no public detections are provided with the dataset, an own detector is trained on the train split of VisDrone MOT. A Cascade R-CNN [5] is used, as the drone images comprise a lot of small objects where this network has its strengths performing the bounding box regression several times with increasing accuracy requirements during training. To adapt the Cascade R-CNN to the dataset, the training is performed on patches of 600 $\times$ 600 pixels and the default anchor sizes are halved to account for the small object sizes. Similarly, to consider the larger number of objects in one image, the number of proposals is doubled. To further

Table 3.1: Results of the Cascade R-CNN detector with different test-strategies.

Cascade R-CNN	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_l
Baseline	35.4	62.5	35.2	12.3	38.9	54.7
+ more proposals	35.8	63.6	35.4	12.5	39.2	54.8
+ multi-scale testing	38.4	67.7	37.6	16.7	42.4	55.7
+ horizontal flipping	**39.2**	**68.5**	**38.4**	**17.8**	**43.2**	**56.6**

improve the detection performance, multi-scale testing and horizontal flipping are used. The influence of these strategies is evaluated on the test-dev split of VisDrone MOT and the resulting average precisions (APs) are summarized in Table 3.1.

3.3 Tracking Results

For a fair comparison, all evaluated tracking methods use the same set of detections generated by the Cascade R-CNN detector with all test-time improvements (see Table 3.1). For Tracktor++, the default Faster R-CNN is exchanged with the trained Cascade R-CNN to use the superior detector also for the proposal regression that implicitly performs the association. Similarly, for all methods, the same re-identification model from [11] and CMC model from [6] are taken, if applicable. The tracking results on the test-dev split are shown in Table 3.2. Note that the AP for MOT differs from the AP for object detection.

The PAS tracker outperforms the other methods by a large margin for both short- ($AP_{0.25}$), middle- ($AP_{0.5}$) and long-term tracking ($AP_{0.75}$) as well as for all object categories. The V-IOU tracker performs the worst, since it only uses IoU for the association. The IoU is not as accurate as the position and the size similarity of the PAS tracker, especially in crowded scenes with small object sizes. This is reflected in the very low AP_{ped} value for pedestrian tracking, as in the VisDrone MOT dataset pedestrians often appear in groups. The Deep SORT algorithm does not rely on IoU but bases its association solely on appearance similarity. However, the extraction of appearance features is harmed by nearby

Table 3.2: Comparison of the PAS tracker with other tracking approaches from the literature. Note that all methods use the same object detector.

Tracker	AP	$AP_{0.25}$	$AP_{0.5}$	$AP_{0.75}$	AP_{car}	AP_{bus}	AP_{trk}	AP_{ped}	AP_{van}
V-IOU	26.4	34.5	29.2	15.6	40.9	36.4	22.7	7.8	24.4
Deep SORT	33.2	51.0	35.0	13.6	31.9	58.3	30.0	21.3	24.5
Tracktor++	34.3	48.6	35.5	18.8	50.6	40.0	32.8	20.2	27.8
PAS	**50.8**	**66.1**	**52.5**	**33.8**	**62.7**	**81.2**	**43.9**	**30.3**	**35.9**

overlapping objects under occlusion and no precise position information is available in the association. In an ablative experiment, it is found that the precise position similarity has the most impact on the tracking performance in the PAS tracker. Whereas Tracktor++ achieves state-of-the-art results on other tracking benchmarks, it struggles in the VisDrone MOT dataset, mainly at small objects and in crowded scenes, since the bounding box regression is sensitive to jumping onto nearby objects. Using position, appearance and size information at the same time, the PAS tracker achieves state-of-the art performance on the VisDrone MOT dataset.

4 Conclusion

In this report, three popular MOT methods originally developed for person tracking are extended to multi-category trackers and tested on a dataset of drone-based imagery. For this, a Cascade R-CNN detector is adapted to the drone images to improve detection performance. After a detailed comparison of the existing trackers, it is found that none of them takes full advantage of object cues and a new tracker that uses position, appearance and size information at the same time in the association step is designed. The proposed PAS tracker outperforms the other approaches by a large margin. In future works, other combination possibilities of the available object information should be investigated.

References

[1] P. Bergmann, T. Meinhardt, and L. Leal-Taixé. “Tracking Without Bells and Whistles”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

[2] A. Bewley et al. “Simple Online and Realtime Tracking”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016.

[3] E. Bochinski, V. Eiselein, and T. Sikora. “High-Speed Tracking-by-Detection Without Using Image Information”. In: *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017.

[4] E. Bochinski, T. Senst, and T. Sikora. “Extending IOU Based Multi-Object Tracking by Visual Information”. In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2018.

[5] Z. Cai and N. Vasconcelos. “Cascade R-CNN: Delving Into High Quality Object Detection”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.

[6] G. D. Evangelidis and E. Z. Psarakis. “Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.10 (2008).

[7] R. Girdhar et al. “Detect-and-Track: Efficient Pose Estimation in Videos”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[8] J. F. Henriques et al. “High-Speed Tracking with Kernelized Correlation Filters”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2015).

[9] H. W. Kuhn and B. Yaw. “The Hungarian Method for the Assignment Problem”. In: *Naval Research Logistics Quarterly* (1955).

[10] L. Lan et al. “Interacting Tracklets for Multi-Object Tracking”. In: *IEEE Transactions on Image Processing* 27.9 (2018).

[11] H. Luo et al. “Bag of Tricks and a Strong Baseline for Deep Person Re-Identification”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.

[12] S. Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017).

[13] D. Stadler, L. W. Sommer, and J. Beyerer. “PAS Tracker: Position-, Appearance- and Size-Aware Multi-object Tracking in Drone Videos”. In: *Computer Vision - ECCV 2020 Workshops*. 2020.

[14] N. Wojke, A. Bewley, and D. Paulus. “Simple Online and Realtime Tracking with a Deep Association Metric”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017.

[15] P. Zhu et al. “Vision Meets Drones: Past, Present and Future”. In: *arXiv preprint arXiv:2001.06303* (2020).

Classifying Usage Control and Data Provenance Architectures

Paul Georg Wagner

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

Abstract

Given the ubiquity of data acquisition and processing in our everyday life, protecting data sovereignty in distributed systems is a significant topic of research. Usage control and provenance tracking systems are very promising steps towards a technical solution for the problem of data sovereignty. However, due to their complexity and diversity these systems are still not fully understood. In this work we investigate the functionality of usage control and provenance tracking systems. We classify them into three different categories based on their security goals and properties. Furthermore we identify generic use cases for these systems that help to understand what attack vectors system operators have to be mindful of.

1 Introduction

In the age of ubiquitous computing, data are quickly becoming the most important assets of many private enterprises and public IT infrastructures. Therefore securely managing databases and preventing cyber attacks as well as data theft have been crucial IT security tasks for quite some time. However, in recent years

the focus of these IT security goals somewhat changed. While in the past it was sufficient to protect local infrastructures such as computer systems and databases from unauthorized access, many modern business processes require extensive data exchange with remote stakeholders such as clients, business associates and customers. Examples for this can be found in the context of digital supply chains and collaborative predictive maintenance [5]. Current research projects such as the *International Data Space* [7] push for highly interconnected business ecosystems on a big scale and in many different areas. In such scenarios valuable business data are being disclosed into computer systems operated by external stakeholders, who might have conflicting interests. From an IT security perspective the data owner needs a way to control his information even when it is being processed in remote infrastructures. It needs to be ensured that the data recipients cannot inadvertently disclose the received information, or even deliberately misuse it for their own benefit.

Similar challenges also exist when considering the topic of data privacy protection. Unlike with business data, personally identifiable information of a single individual seldom holds great monetary value. Nevertheless the protection of shared personal information is still of great concern. While (supra-)national data privacy laws regulate the acquisition and usage of personally identifiable information on the legislative level, given the noticeable trend towards highly interconnected data processing systems, there is a clear demand for technical solutions as well. At the present moment this is especially evident in the field of medical data processing. In light of the current global Covid-19 health crisis, being able to autonomously collect and distribute health data on a large scale has become profoundly relevant. Nonetheless, given the privacy-sensitive nature of these data, the patients clearly need to remain in control of their information throughout this process. As a result, over the last few years lots of research regarding privacy-compliant medical data sharing has been conducted [1, 3, 6]. In general, data subjects have a legitimate right to monitor and control what personal information is being used in what way, even if the actual data processing is performed on a remote device operated by a third party.

These challenges regarding both data protection and data privacy can be subsumed under the term *data sovereignty*. Data sovereignty describes the approach of enabling data providers to monitor and control the use of their

information at all times, even when they are being used by remote stakeholders. During this process, the data in question can be business-related, or consist of personally identifiable information. To achieve data sovereignty in technical systems, there are some tasks to be considered. First of all, it is necessary to track data flows across systems and domain boundaries. This is called *data provenance tracking* and allows to reliably monitor data usage regardless of where the data is being processed. Besides passively observing data flows, providers also need to be able to actively control and prevent certain types of unwanted data usage. This can be done by applying *usage control* (UC) techniques.

In this work we investigate the general design of usage control and provenance tracking systems and analyze them with regard to four dimensions:

- Security goals
- Enforcement capabilities
- System architecture
- Attack vectors

Based on this analysis we classify usage control and provenance tracking systems into three different categories. We also identify generic use cases for these systems that help to understand what stakeholders are relevant and what attack vectors can occur in different scenarios. The remainder of this paper is structured as follows. Section 2 briefly introduces the design and functionality of usage control and provenance systems on a purely conceptual level. Afterwards in section 3 we identify and categorize several corresponding system architectures that are used as a basis for implementing these concepts. In section 4 we then identify relevant stakeholders and describe four generic use cases for usage control and provenance systems as well as important attack vectors that have to be considered. Finally in section 5 we conclude with a short recap and an outlook on future research.

2 Provenance Tracking and Usage Control

In order to establish a technical solution for data sovereignty, we need versatile provenance tracking and effective usage control frameworks. Both of these topics have been subject to a lot of research in the past. In this section we will briefly introduce the most common way of defining provenance tracking and usage control mechanisms.

2.1 Provenance Tracking Mechanisms

Data provenance allows data providers to track the usage of their digital assets and collect information about derivations that have been created as part of a data processing step. The most common formal model for provenance is the PROV standard [4], formerly known as the Open Provenance Model (OPM). This family of documents describes data formats and serializations for exchanging provenance information across heterogeneous environments. It does not, however, propose concrete mechanisms for implementing provenance tracking in data processing systems. For this, Bier [2] suggests a provenance tracking system mainly consisting of three distinct components (c.f. figure 2.1).

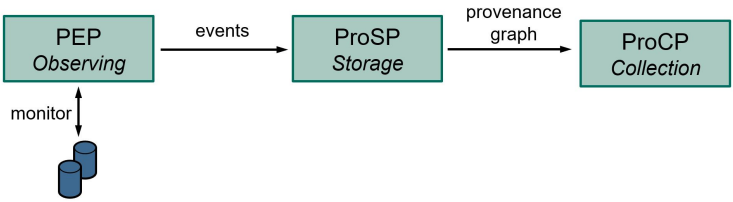


Figure 2.1: Provenance tracking components.

First of all, a *policy enforcement point (PEP)* is responsible for monitoring data accesses and creating events that represent data flows within the system. PEPs are usually implemented close to data processing applications and are capable of examining how data is being used. As data monitoring components, they are at the heart of each provenance tracking system. The generated events

containing data flow information are then relayed to a *provenance storage point (ProSP)*. The ProSP evaluates the events and aggregates all data flow information into a provenance graph. The nodes of this provenance graph correspond to representations of certain data at a specific point in time, while the edges describe linkages between data representations (i.e. data flows). In short, the provenance graph represents a comprehensive information flow history of the entire data processing domain. If sensitive data are shared across domain boundaries, a third level is established by including a *provenance collection point (ProCP)*. This component queries and aggregates multiple provenance graphs, thereby creating a coherent history of data that have been tracked across multiple systems.

2.2 Usage Control Mechanisms

In addition to tracking provenance information, achieving data sovereignty requires a mechanism for data providers to actively and continuously control the access to their information even after it has been disclosed. This can be done by applying usage control (UC) techniques. Usage control was developed over a decade ago as a generalization of attribute-based access control. In contrast to classical access control schemes, UC allows for continuous authorization of data accesses over a period of time. It also features the possibility to declare obligations that need to be fulfilled before, during or after a certain data usage, which is not covered by classical access control. This allows the definition of complex data usage strategies, such as limiting the number of views or the time of access to sensitive information. The most widely adopted formal usage control model is $UCON_{ABC}$, which has been introduced in 2004 by Park and Sandhu [8]. Even today $UCON_{ABC}$ provides the formal basis for many usage control systems. In terms of designing usage control architectures, most modern systems rely on a derivative of the XACML reference architecture [9]. Originally developed for attribute-based access control, the XACML components can be canonically extended to implement usage control policies. Figure 2.2 shows a generic usage control system based on XACML components.

As before, the central component of the system is a policy enforcement point (PEP), which closely interfaces the data processing applications and continuously generates events representing any data usage. However, unlike PEPs

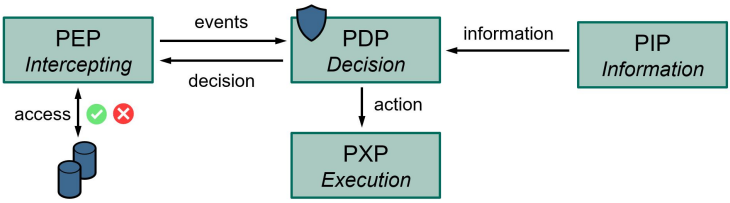


Figure 2.2: Usage control components.

implemented for provenance tracking, usage control PEPs must be capable of actively interfering with the data processing as well. It is not sufficient to just *observe* data usage anymore – usage control PEPs need to be able to actively *intercept* data usage events and potentially *modify* or *block* their execution based on the prevailing usage policies. We call an enforcement point capable of this *intercepting PEP*, in contrast to an *observing PEP*. Naturally every intercepting PEP is also an observing PEP. While observing PEPs are sufficient for provenance tracking, we require the definition of suitable intercepting PEPs in order to enforce usage control on sensitive data.

The other essential usage control component is the *policy decision point (PDP)*. The PDP holds a set of active usage control policies and receives the events from the intercepting PEP. The received events are then evaluated against the set of active policies, which results in a usage control decision. In addition to the classical binary access decision of `allow` versus `deny`, the PDP can also rule that the data usage described by the event should be `modified` prior to its execution. In the end the intercepting PEP receives the decision and enforces it on the data processing application.

Finally there are two more components involved in the usage control enforcement process. The *policy information point (PIP)* can be queried by the PDP for subject and object attributes, as well as generic information such as database entries or environmental properties. The *policy execution point (PXP)* is responsible for executing obligations demanded prior to a data usage, for example the incrementation of an access counter. Obligations are invoked by the PDP and have to be executed successfully before the PDP publishes a positive decision.

In the end it is the collaboration of all components that ensures proper usage control enforcement.

3 Classifying System Architectures

In the previous section we described the mechanisms and basic components of usage control and provenance tracking systems. However, this merely conceptual view on usage control and provenance does not yet describe how to actually apply these mechanisms in real-world use cases. Depending on the specific demands and requirements there are many ways of designing usage control and provenance architectures. In this section we explore and categorize different options in designing actual system architectures and discuss what real-world use cases they cover.

3.1 Usage Control Architectures

Local UC architecture. The simplest form of usage control systems are *local UC architectures*. Figure 3.1 shows an example of such an architecture. A local UC architecture consists of a single computer system that enforces a set of usage rules on local data without considering any external influences. The usage control components (PEP, PDP, PIP and PXP) all run as services on the local computer system that should be protected. Furthermore, both the sensitive data as well as the respective usage control policies are also stored directly on this system. During system operations, the usage rules are then enforced on the local database by the mechanisms described earlier (c.f. figure 2.2). Usually there is a fixed set of usage control policies that have been defined by the system administrator (analogous to mandatory access control). In addition to that, system users can also create protection policies for their own data (analogous to discretionary access control).

Figure 3.1 shows the four usage control components implemented as dedicated software modules instead of a single large monolithic software stack. Usually this design is preferred, because it respects separation of concerns and allows the

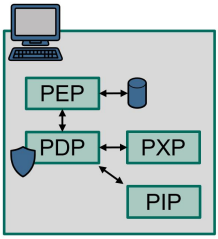


Figure 3.1: Local usage control architecture.

flexible extension of the usage control system. Nevertheless, all UC components run on a single system containing all data that are subjected to usage control.

The benefit of the local UC architecture is its simplicity and self-containment. The architecture does not depend on any other system and manages both the database as well as the policy set sovereignly. On the other side, the local UC architecture cannot enforce usage control anymore as soon as the data are being shared with another system. Hence it is not suitable for implementing any of the scenarios that have been described in the introductory motivation.

Cross-domain UC architecture. In order to support usage control enforcement across different stakeholders, multiple local UC architectures can be merged into a *cross-domain UC architecture*. As the example in figure 3.2 shows, a cross-domain architecture links together multiple remote UC systems that can share sensitive data as well as respective usage control policies. Each participating usage control system represents a single UC domain, i.e. it operates on a set of policies that are evaluated by a single decision point (PDP). Even though the cross-domain architecture now deals with multiple UC domains (unlike local architectures), each UC domain is still implemented as a single computer system.

The main difference of this architecture compared to a set of local UC architectures is that now data flows between usage control domains are being considered as well. Furthermore, the various UC domains are usually operated by different stakeholders. For example, figure 3.2 shows a data flow from the green system on

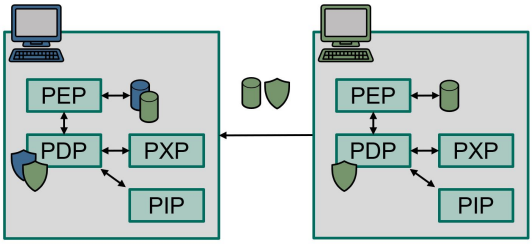


Figure 3.2: Cross-domain usage control architecture.

the right to the blue system on the left. Since the data provider wants to protect his sensitive information in the domain of the remote data receiver, the data flow is preceded by the deployment of a usage control policy regulating the data usage on the remote side. This deployment step is automatically initiated by the enforcement point (PEP) of the donating system (here: right side) whenever it observes an outgoing data flow. The actual policy transmission is then performed by the policy execution point (PXP) as part of a UC obligation. The donating enforcement point only allows the outgoing data flow if the deployment of the protection policy has been executed successfully. Finally the deployed policy is being continuously evaluated by the remote decision point on the data receiver (here: left side). At this point the remote PEPs ensure proper enforcement of the demanded usage restrictions even outside the data owner's usage control domain.

Being able to handle usage control between different stakeholders is the main advantage of cross-domain UC architectures over purely local architectures. Hence cross-domain architectures are suitable to implement the scenarios outlined in the introductions. On the other side, now the usage control systems must be able to remotely deploy and enforce protection policies. Furthermore, the used policy model has to be able to distinguish local from remote data usage. Finally the cross-domain architecture is still limited to a single computer system per usage control domain. This hinders scalability in generic and flexible use cases.

Distributed UC architecture. The most flexible type of usage control systems are designed as *distributed UC architectures*. In contrast to the cross-domain configuration, distributed architectures allow the deployment of a single usage control domain over several collaborating computer systems. Figure 3.3 shows an example of this type of usage control. As you can see, the usage control components previously running on a single computer system are now distributed across multiple devices. Most notably, there are now multiple PEPs running on user devices, while the policy decision point (PDP) runs centrally on a dedicated server. This allows support for use cases where several computer nodes are used for data processing inside a UC domain (e.g. when using thin clients in server environments). Depending on the scenario it is also possible to include multiple information points (PIPs) and execution points (PXP) running on dedicated hardware. This is very useful, since both of these components often attach to existing infrastructure such as databases or directory services. However, in order to avoid policy conflicts there is usually only a single decision point (PDP) per usage control domain.

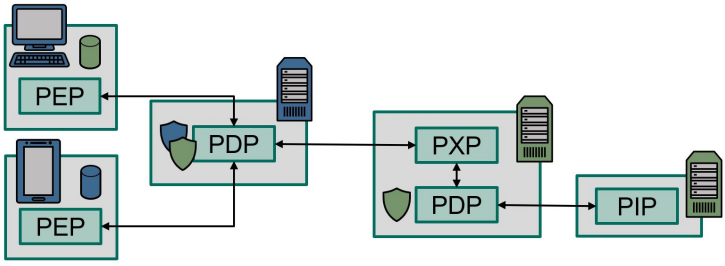


Figure 3.3: Distributed usage control architecture.

While in principal suitable for the same usage control scenarios as cross-domain architectures, the distributed form of usage control offers far greater flexibility than their monolithic counterparts. This is true even in scenarios with only a single UC domain and no data flows between different stakeholders. Being able to independently deploy PIPs and PXPs as dedicated components in existing server infrastructure enables a wider range of usage control applications. Supporting multiple PEPs within a single UC domain allows data processing applications

to be deployed independently of the rest of the usage control infrastructure, for example on mobile devices of employees. Furthermore, this also offers a level of scalability within a UC domain. It is now possible to add more enforcement, execution or information points into an existing UC domain whenever the need arises. On the other hand a distributed UC architecture is more complicated than a UC system running exclusively on a single computer system. It has to be ensured that all usage control components can communicate reliably and securely with each other, even when they are located within a single UC domain. Because of this increase in complexity there are broader attack vectors on distributed usage control architectures and their security properties must be inspected more closely.

3.2 Provenance Tracking Architectures

In addition to distinguishing different types of usage control architectures, provenance tracking systems can be classified in a similar fashion. However, unlike the UC architectures, provenance tracking systems should be classified according to the scope of the acquired provenance information rather than how the system components are deployed.

System-wide provenance tracking. The simplest way of tracking provenance information is to only focus on the data within a single computer system. As the example in figure 3.4 shows, such a system consists of at least one policy enforcement point (PEP) observing all data usages on the system, while a provenance storage point (ProSP) residing on the same system generates and stores provenance information. While there is always only one ProSP per system, it is possible to use multiple PEPs for monitoring data usage (e.g. one per data processing application). The generated provenance graph then attests to the history of data usage on this particular system, for example in order to prove compliance with data privacy laws.

Naturally, this type of architecture only tracks the provenance of data while it is being processed on a single computer system. As soon as the information leaves the system in question, no more fine-grained provenance tracking is

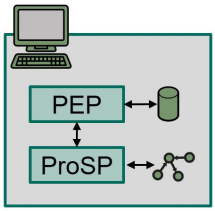


Figure 3.4: System-wide provenance tracking architecture.

possible. This is because the provenance storage points have only a local view on data processing and operate independently on multiple systems. Hence this architecture is only suitable for use cases where all relevant data processing is performed inside a single system.

Domain-wide provenance tracking. Domain-wide provenance tracking is used to track the provenance of data within a single domain. Unlike the system-wide provenance tracking, this means that data flows between multiple systems (i.e. multiple PEPs) within a domain are being tracked by a dedicated ProSP. However, it is still not possible to track data flows across multiple domains and different stakeholders. Figure 3.5 shows an example of this architecture.

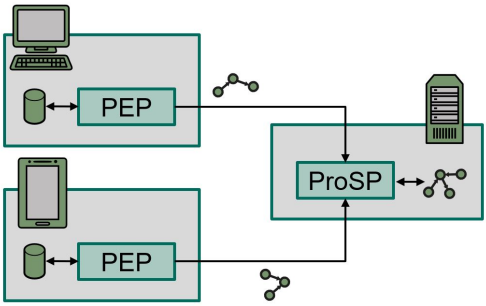


Figure 3.5: Domain-wide provenance tracking architecture.

Cross-domain provenance tracking. In contrast to the previous two architectures, *cross-domain provenance tracking* allows keeping track of data usage even across multiple domains and stakeholders. For this, multiple domain-wide (or system-wide) architectures are linked together and share provenance information. That way provenance tracking is possible even across the domains of different stakeholders. In addition, there is a global provenance collection point (ProCP) aggregating the provenance graphs of multiple local provenance storage points. This allows to generate a comprehensive history of data usage across multiple domains. Figure 3.6 shows an example of cross-domain provenance tracking.

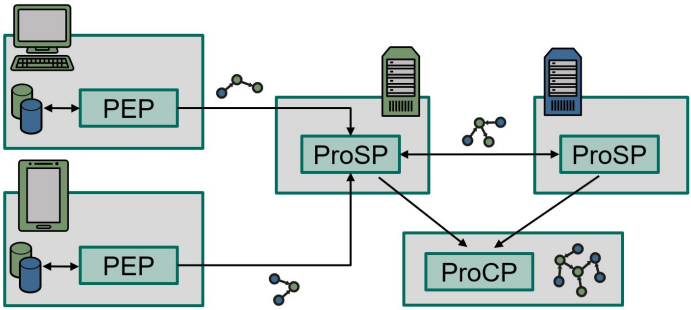


Figure 3.6: Cross-domain provenance tracking architecture.

As Bier pointed out in [2], usage control and provenance tracking can be combined. Similarly, usage control and provenance architectures can also be combined. Clearly, the system-wide provenance tracking architecture is compatible with both local and cross-domain UC architectures. Since there is at least one policy enforcement point in each of those architectures, all that is needed to implement system-wide provenance tracking is a provenance storage point for each system. However, system-wide provenance tracking cannot be used with distributed architectures, since there is no single computer system performing all relevant data processing anymore. On the other hand, domain-wide provenance tracking requires a distributed usage control architecture, while cross-domain provenance tracking is compatible with cross-domain and distributed usage control. Both of them are not compatible with purely local usage control,

because their PEPs cannot observe data flows across system and/or domain boundaries. Table 3.1 shows the possibilities of combining usage control and provenance architectures.

Table 3.1: Combining usage control and provenance architectures.

		Usage control		
		local	cross-domain	distributed
Provenance	system-wide	✓	✓	✗
	domain-wide	✗	✗	✓
	cross-domain	✗	✓	✓

4 Identifying Generic Use Cases

After describing the different possibilities of realizing usage control and provenance architectures, we are left to evaluate their security properties in different use cases. We do this by first identifying the stakeholders that have an interest in usage control and provenance systems of different flavors. As we will see, depending on the concrete goal of the protection systems, the stakeholders' motives can change somewhat and they even have to be considered attackers. Afterwards we describe four different generic use cases for provenance and usage control that demonstrate what attack vectors are to be expected and what security guarantees the various options ultimately yield.

4.1 Stakeholders

There are four main stakeholders to be considered when designing usage control and provenance architectures.

Data owner. This stakeholder holds the rights on a certain set of data that is being disclosed. Usually the data owner has either a monetary or personal interest in monitoring and regulating the usage of his information. For this

purpose the data owner defines usage control policies that specify what may or may not be done with the disclosed information. Furthermore, the data owner may demand tracking the provenance of his information to maintain transparency. If the disclosed data contain personally identifiable information, a data owner is often also called *data subject*.

System user. A system user operates computer systems that run data processing applications in the usage control infrastructure. This stakeholder has legitimate access to information previously disclosed by a data owner and uses it to achieve a certain task. While doing so, the usage control infrastructure enforces the restrictions provided by the original data owner. The system user's access and distribution of protected information may also be logged by the provenance tracking infrastructure. Crucially, the system user does not have privileged access to the systems he operates or the protection components running there. Depending on the scenario, a system user may be motivated to bypass the usage control protection and/or provenance tracking for his personal benefit. Hence this stakeholder must be considered as a possible adversary.

System owner. The system owner is responsible for operating computing systems that run data processing applications as well as the local usage control and provenance infrastructure. Usually this stakeholder has an interest in receiving sensitive information from external data owners and use them outside the boundaries specified by the data owner's usage rules. Unlike the system user he has privileged access to all parts of the managed infrastructure and may use this power to manipulate usage control and provenance components. However, as we will see there are also scenarios where the system owner is the primary data owner. In this case we can trust the system owner to setup all protection systems correctly and not bypass the usage control enforcement, since it is his own interest to enforce protection rules against non-privileged system users.

Supervisory authority. The supervisory authority is a stakeholder only relevant to infrastructures with provenance tracking. This stakeholder is not directly involved in any data sharing, but instead has an interest in verifying

the legitimacy of the data usage with regard to data privacy laws. Usually this stakeholder is a government agency, but it may also be a trusted third-party verifying the privacy-compliance for all participants.

4.2 Provenance Compliance

Provenance compliance is a use case where stakeholders want to track the usage of sensitive information throughout the whole data processing infrastructure. Their goal is to verify the legitimacy of the data usage and its compliance to contractual agreements or data privacy laws. We can distinguish between *internal compliance* and *external compliance*.

Internal compliance means that a system owner intends to conduct provenance tracking only within his own local infrastructure and on his own data. An example for this use case would be a company tracking data flows within their own infrastructure for process optimization purposes or to verify that their employees act in compliance with company-internal standard operating procedures. In this case the system owner (i.e. the company) simultaneously acts as the data owner and the supervisory authority. Notably there are no external data owners present in this scenario. The only other relevant stakeholders are the system users, which in this example would be company employees using the IT infrastructure to perform their tasks as usual, thereby generating provenance information. While the company acts as supervisory authority by analyzing the provenance on their own data, the generated provenance graphs are not intended to serve as evidence for any external supervisory authority. For the use case of internal compliance both system-wide and domain-wide provenance architectures are suitable. Regarding the security properties of internal compliance, the most important adversaries are the system users (i.e. the employees). They might try to hide illegitimate or undesired data usages from their employer by blocking or forging provenance information. Hence the system owner needs to make sure that the provenance tracking components (mainly PEP and ProSP) are properly protected from tampering. In case of a domain-wide provenance architecture it also has to be ensured that no forged provenance information can be sent to the provenance storage point. However, even though a system user may be

motivated to tamper with provenance information in order to hide data accesses from his employer, there is usually no direct monetary incentive for him to do so.

External compliance on the other hand is conducted with the explicit goal of proving compliance to external data owners or supervisory authorities. In this case the system owner receives sensitive information from external data owners in order to process it in his own infrastructure. He is required to track the usage of this data in his domain and report the respective provenance information back to the data owners for transparency. All three discussed provenance tracking architectures are suitable for this task. The most important difference to internal compliance is that the system owner now has a clear interest of forging provenance information in order to deceive the original data owners and supervisory authorities. He can do this by manipulating either the enforcement points or the provenance storage points on his own systems. In order to mitigate this problem, we need to establish trust in the system owner's infrastructure, either by contractual agreements or technical measures (c.f. section 5).

4.3 Usage Control Enforcement

Besides provenance compliance, the other important use case is enforcing usage control on data processing applications. Depending on the goals that should be achieved by the protection system, once again we can distinguish between *internal enforcement* and *external enforcement*.

Internal usage control enforcement, similar to internal compliance, is conducted by a system owner solely in his own infrastructure. For example, a company has a valuable pool of business data and wants to safeguard data usage in their own local infrastructure. The system owner can do this by establishing either a local or distributed UC architecture (in the latter case only with a single domain). In such a system intercepting enforcement points will oversee all data usages in the companies infrastructure and query the central decision point for each data access. The company can then deploy proper usage restrictions in form of policies at this decision point. Once again, since in this case the system owner is simultaneously also the data owner, we can trust him to properly setup and operate the necessary usage control infrastructure. However, the system users (i.e. employees) that are being subjected to usage control enforcement when

working with the protected data, may be motivated to bypass the protection components and access the sensitive data without restriction. Hence we have to view system users as the most important possible adversaries and properly protect all the deployed usage control components from tampering by non-privileged system users.

External usage control enforcement is the most important use case that a UC architecture should address. In this case a data owner wants to impose his own usage restrictions on *remote* data processing applications running within a remote UC domain. This requires either a cross-domain or distributed usage control architecture. The data owner can then deploy his own usage control policies into the remote UC domain, before transmitting sensitive data. The remote usage control components (either within a single computer system or as distributed services) then intercept all data usages in the remote system and enforce the deployed usage restrictions on them. As before, from a security perspective the most important adversary is the system owner of the receiving side. This remote system owner has a clear interest of bypassing the usage control components in his own domain and access the foreign data without restrictions. Furthermore the system owner has full control over the enforcing UC system and can manipulate the components to either ignore the deployed policies, or bypass the interception at the enforcement point. Just as with external compliance, we hence need to establish trust in the system owner's *remote* infrastructure, either by contractual agreements or technical measures.

Table 4.1 shows the four identified use cases and their characteristics.

Table 4.1: Use cases for usage control and provenance.

	Compliance		**UC-Enforcement**	
	internal	**external**	**internal**	**external**
PEP capability	observing	observing	intercepting	intercepting
Architecture	system domain	system domain cross-dom.	local distributed	cross-domain distributed
Attacker	user	owner	user	owner

5 Conclusion

In this work we investigated the functionality and properties of usage control as well as provenance tracking systems. We classified three different variants of both UC and provenance tracking architectures, and described in what way they can be combined. Finally we identified generic use cases of usage control and provenance tracking systems that help to understand how these systems can operate in practice and what attack vectors are to be expected.

As pointed out in the previous section, there has to be a way of enforcing the correct application of these techniques on the remote side. This is especially important for the use cases of external compliance as well as external usage control enforcement. This part of the problem is often overlooked, but it is essential for the security guarantees of the resulting system. While most existing systems make do with non-technical solutions such as operational and contractual agreements, a possible technical solution for this issue relies on *trusted computing* to cryptographically attest to the integrity of usage control and provenance tracking components. However, correctly applying trusted computing to this problem is not trivial and still an area of active research [10]. Since there are many ways of ending up with insecure systems when not properly considering how to establish trust in remote usage control and provenance components, this is an important area for future research.

References

[1] Arno Appenzeller et al. “Enabling data sovereignty for patients through digital consent enforcement”. In: *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. 2020, pp. 1–4.

[2] Christoph Bier. “How usage control and provenance tracking get together-a data protection perspective”. In: *2013 IEEE Security and Privacy Workshops*. IEEE. 2013, pp. 13–17.

[3] Eleonora Ciceri et al. “PAPAYA: A Platform for Privacy Preserving Data Analytics”. In: *Digital Health* (2019), p. 42.

[4] Paul Groth and Luc Moreau. “PROV-overview. An overview of the PROV family of documents”. In: (2013).

[5] Matthias Jarke, Boris Otto, and Sudha Ram. *Data Sovereignty and Data Space Ecosystems*. 2019.

[6] Xiaoguang Liu et al. “A blockchain-based medical data sharing and protection scheme”. In: *IEEE Access* 7 (2019), pp. 118943–118953.

[7] Boris Otto et al. *IDS Reference Architecture Model*. Tech. rep. International Data Spaces Association, 2018.

[8] Jaehong Park and Ravi Sandhu. “The UCON ABC usage control model”. In: *ACM Transactions on Information and System Security (TISSEC)* 7.1 (2004), pp. 128–174.

[9] OASIS Standard. *extensible access control markup language (xacml) version 3.0*. 2013.

[10] Paul Georg Wagner, Pascal Birnstill, and Jürgen Beyerer. “Challenges of Using Trusted Computing for Collaborative Data Processing”. In: *International Workshop on Security and Trust Management*. Springer. 2019, pp. 107–123.

Learning Universal Vector Representation for Objects of Different 3D Euclidean formats

Chengzhi Wu

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

Abstract

We present a method for learning universal vector representations out of 3D objects represented in different data formats. A newly proposed switching mechanism is used in the design of neural network architecture. During the learning process, the encoder for one specific format also learns to perceive the object from the perspective of other formats, hence the learned universal representation contains richer information. With the learned universal representation, it would be possible to "translate" between different 3D shape formats of the input object since they share similar embedding of 3D information. Higher performance can also be achieved for the 3D data synthetic tasks with this method.

1 Introduction

1.1 Latent representation of 3D data

Depending on the measuring method and the processing and storing rules of information, 3D objects may have various representing formats in the real-world. On the Euclidean side, they may be represented as RGB-D images, multi-view images or volumetric data. On the Non-Euclidean side, they may be represented

as point clouds or meshes. However, no matter in which format store the 3D information of the object, when it comes to the computer vision tasks, e.g. detection, segmentation, or even other generative tasks, the target 3D object will usually need to be converted into a latent representation first for further computation.

Before the surge of deep learning, it was common to use classical mathematical algorithms to get those 3D shape latent representations (or, 3D shape descriptors). This computation process usually involves strict mathematical formulas and deductions to get rule-based representations, e.g. Laplacian spectral eigenvectors [15], or heat kernel signature [17]. Thanks to the development of deep learning algorithms, the performance of some computer vision tasks, especially in the detection and segmentation domain [7], have been boosted. During the training of those neural networks, latent representations of input have already been generated implicitly. Although this learning process has been regarded as a black box at earlier years, researches in the visualization of learned latent representations have been conducted [24]. Throughout the computer vision learning history, a better method for learning the latent representations leads to better performance on those tasks.

1.2 Universal vector representation

Learning an universal vector representation touches on two long-standing and important questions in computer vision: how do we represent 3D objects in a vector space and how do we recognize this representation from images. [6] believed that a good vector representation for objects must satisfy two criteria: it must be (1) generative in 3D; (2) predictable from 2D. In this report, we learn universal representations for 3D objects of different formats by leveraging the advantages of deep learning algorithms. For simplicity, we are only investigating Euclidean data in this report.

On the one hand, the vector representation can be learned from different data formats such as multi-view images and volumetric data; On the other hand, it can be inferred during the training process of different neural networks designed for different machine vision tasks. For analytical tasks, especially for the classic classification tasks, vector representations will usually be learned before the

last several fully connected layers. These vector representations are sometimes referred as bottleneck features. Those bottleneck features can be further adapted for other tasks, as it is done in transfer learning. For synthetic tasks, typical generative models are AE/VAE [11] and GAN [8]. They learn the mappings between the latent space and the real-world data space, thus reconstructions from latent representations are possible. Theoretically, if we can learn universal representations that contain both view information and geometry information, better synthetic results may be achieved. Hence those generative models may be modified for learning universal representations in our case and may also be used as a verifier to indicate the performance.

From another perspective, the process may also be regarded as data compression and the richness of its implicitly stored feature information is of pivotal importance. Since the learned universal vector representations can be used not only for synthetic tasks, but also for analytical tasks, we are also expecting higher performance in regular machine vision tasks like classification with them. Since we want to merge the information from different data formats, the resolution of data should also be considered. It would be apparently inappropriate to have a fixed-size latent vector to represent objects of different resolutions, even under the same format. Hence, the main idea of this report is to learn a fixed length of vector representation for an object of a specific category, under certain resolution limitations of different data formats.

2 Related Work

2.1 Learning representations (encoders)

Although latent representations are also learned in analytical tasks, they have been seldom specifically explored. There are numerous papers using various network architectures for 3D machine vision tasks. A typical one is VoxNet[13], which was the first to use 3D convolution operations to learn features from volumetric data. Its subsequent work of multi-level 3D CNN [5] learns multi-scale spatial features by considering multiple resolutions of the voxel input. Regarding the multi-view images format, a typical method is MVCNN [16]. It

uses a parameter sharing network to encode images of one object from different views, followed by a view pooling layer before the last several fully connected layers. A subsequent work of GVCNN [4] groups all the views before encoding. Each group uses one separate parameter sharing network to encode this group of images, then a group fusion operation is defined in the latter step.

Methods combing the information from both multi-view images and volumetric data have also been proposed. For example, Qi et al. [14] proposed to use multi-resolution filtering in 3D for multi-view CNNs, as well as using subvolume supervision for auxiliary training. Another example is FusionNet [10], which is a fusion of three different networks: two VoxNets and one MVCNN. The three networks fuse at the score layers where a linear combination of scores is taken before finding the class prediction. Voxelized CAD models are used for the first two networks and 2D projections are used for latter network.

2.2 Generating from representations (decoders)

Unlike analytical tasks, latent representation matters a lot to synthetic tasks. There are mainly two types of deep generative models nowadays: AE/VAE [11] and GAN [8]. Based on those two frameworks, various methods have been proposed to learn latent representations from 3D data and to reconstruct back to them. ShapeNet [23] used a reverse VoxNet, i.e. a decoder, to reconstruct 3D shapes from a latent representation which was learned from depth maps. The dataset they created is also being widely used for 3D machine vision tasks nowadays. Girdhar et al. [6] used AE directly to encode and decode 3D shapes. With the learned latent representation from volumetric data, they proposed a TL-embedding network which forces another encoder to learn a exactly the same latent representation from corresponding images. This makes it possible to generate 3D shapes from images. VAE has also been used in a similar way for the 3D shape learning in other paper [2].

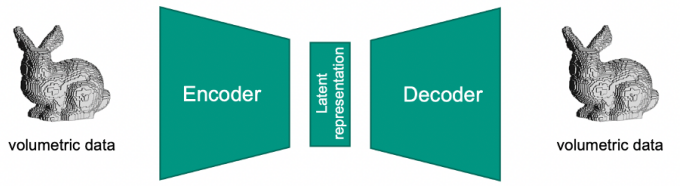
View information from images has also been widely investigated for 3D shape reconstruction. Choy et al. [3] proposed a framework named 3D-R2N2 to reconstruct 3D shapes from single- or multi-view images. By leveraging the power of Long Short-Term Memory(LSTM), they discovered that the reconstruction is incrementally refined as the network sees more views of

the object. Some other papers also have used view information as auxiliary constraints for the training of their 3D AEs. Tulsiani et al. [18] trained an additional pose CNN to add an additional consistency loss between the inferred depth image from a perspective and its ground truth. This inferred ray-trace pooling view has also been used in the adversarial part of [9] for weakly supervision.

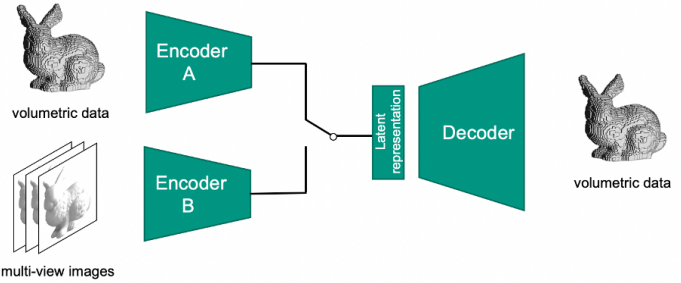
Methods used GAN for 3D shape generation have also been proposed. 3D-GAN [21] makes it possible to generate novel and relatively highly realistic 3D model in the unsupervised way. It introduced three loss functions regarding image encoder, generator and discriminator respectively. The update of all components in the framework is also possible. Besides, visualizing the representation vector, interpolation, arithmetic have been conducted to analyse the vector representations. Liu et al. [12] adapted this idea and proposed an interactive modeling framework that can generate realistic volumetric data with edit and especially defined snap operations. Semantic information has been used in Global-to-Local GAN (G2LGAN) [19] and SAGNet [22] to improve the synthesis quality. G2LGAN also proposed a part refiner to refine the individual semantic part output from local GANs. While showing that segmented information from 3D data can be embedded into the latent space, their work does not include too much discussion of the latent space and its connection with other data formats like image or common voxel.

3 Methodology

Although both AE and GAN were developed for data synthesis tasks by using neural networks, they are quite different in their kernel ideas. AEs use real-world data as input. An encoder-decoder structure network is used to encode the input into latent representation, and subsequently reconstruct it back from the latent representation. The most important loss here is the reconstruction loss. With a well-trained decoder, it is possible to reconstruct the object with a well-learned latent representation. An unsolved question here is how can we force the latent representation to be meaningful. GANs are totally different from AEs since they do not use real-world data as input directly. Instead, they train a generator,



(a) Vanilla Autoencoder



(b) Switch-Autoencoder

Figure 2.1: An illustration of the (a) vanilla Autoencoder (AE) and the proposed (b) Switch-Autoencoder (SAE). AE only takes volumetric data as input, while SAE takes input from both image data and volumetric data, using a switch to randomly choose the learning source. The feature maps/vectors learned inside the network may be regarded as latent representations.

which is similar to the decoder in AEs, on the latent space directly. Generated data will be processed into a discriminator to classify it is generated or from the real world. The whole training process is essentially the competition between the generator and the discriminator. For GANs, the distribution of the latent representation is usually pre-defined as a Gaussian, but how to disentangle the features in the latent space is still a tough question.

In our case, since we are interested in learning a universal representation from 3D data of multi-formats, the original 3D information should be fully utilized. Hence here we adopt the AE architecture to learn latent representations. GAN

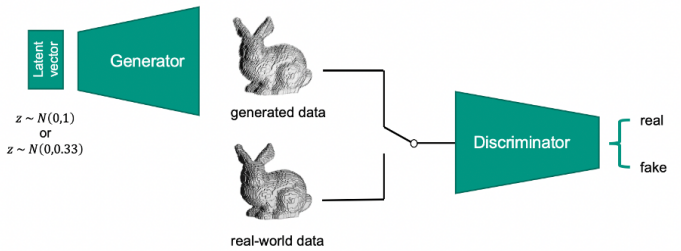


Figure 3.1: The basic structure of a generative adversarial network (GAN). It trains a generator to decode the latent vector representation to a 3D shape, by using a discriminator to force the generator to generate shapes as real as possible.

is great for generating the 3D data, it may be combined with the AE part for better reconstruction in the future step.

3.1 Vanilla Autoencoder

Firstly, we started testing our idea with a simple Autoencoder. As shown in Figure 2.1(a), it is just a normal AE but with 3D convolutions. The loss of this network is the reconstruction loss. There are different ways to compute the reconstruction loss including MSE, Cross Entropy, and IoU. IoU is more like an indicator and does not provide smooth gradient. MSE is mainly used for preliminary tests. In our case, we use the cross entropy as loss function. The output before the last layer has been rectified to a range from 0 to 1. The dataset we are using here is the ShapeNet [23]. It provides a wide variety of real 3D objects, which makes the data-driven learning and analysing of the latent representation possible and promising. The synthesis result from this architecture may be regarded as the baseline of performance.

3.2 The Switch-Autoencoder

The multi-view images data is added to the input side in this setting of experiment. Here, we propose a Switch-Autoencoder (SAE) for universal latent representation

learning. We train two encoders separately for the voxel input and the image sequence input. A switch is attached before the decoder. During the training, the network randomly selects the encoded output from one encoder as the latent representation, then inputs it to the decoder. This operation of switching between encoders continues during the whole end-to-end training. In the TL-embedding network proposed in [6], the image encoder is forced to learn the same embedding from that of the voxel encoder, hence the image encoder does not contribute to the improvement of the generator. Unlike TL-embedding network, in our case, both encoders learn to perceive the object from the perspective of the other format, hence both encoders contribute to the improvement of the generator.

The structure of proposed SAE is shown in Figure 2.1(b). For the image encoder (encoder B), we use an architecture that is similar to multi-view CNN [16]. Each view is encoded with a parameter-sharing network, followed by a view pooling layer. Then it will be passed through several additional fully connected layers to get the final latent vector representation. Here, we also use a network with residual blocks in the image view encoder.

3.3 GAN

In order to improve the synthesis quality, the framework of GAN may be integrated here. In this report, we are focusing on examining the quality of generated shapes from GAN with normal Gaussian vector input. Its basic structure is shown in Figure 3.1. The discussion and experiments of combining AE/VAE and GAN is in the scope of our next step.

There are several non-negligible problems in the vanilla GANs. When training the standard GAN, the loss of the discriminator and generator can oscillate gradually, which make the training process unstable. Besides, vanilla GANs also have the problem of mode collapse, which produces limited varieties of samples. Here, we use the Wasserstein GAN(WGAN) [1] with gradient penality [20], which has two main benefits: (1) improved stability of training process; (2) a meaningful loss metric that correlates with generators convergence and sample equality. We believe that merging the WGAN in the framework will improve the quality of object generation.

4 Experimental Results

4.1 Vanilla Autoencoder

Figure 4.1 shows some results from the vanilla Autoencoder. From it we can tell that the network is able to reconstruct the input 3D shapes from learned latent representations. Besides, the features of different types of chairs have also been well captured.

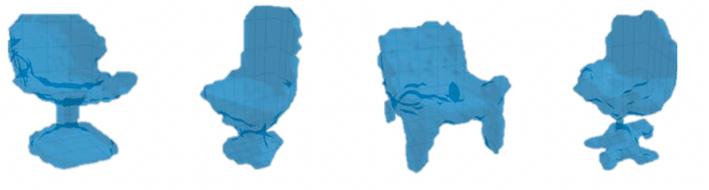


Figure 4.1: Reconstruction results of chair objects from the vanilla Autoencoder.

4.2 The Switch-Autoencoder

Some reconstruction results from SAE is given in Figure 4.2. The up row shows the results with volumetric data as the test input. The bottom row shows the results with multi-view image sequence as the test input. From it we can see that the voxel-encoder still preserves a relatively high quality, while the image-encoder also captures decent 3D information. The sharp areas are difficult for the image-encoder, as can be observed from the generated chair legs.

Overall, comparing with the results from only one format source, objects generated from universal representations with both format sources look better. The generated chairs are usually less rough.

(a) SAE reconstruction results with volumetric data input

(b) SAE reconstruction results with image data input

Figure 4.2: Reconstruction results of chair objects from the proposed Switch-Autoencoder, using (a) volumetric data or (b) image data as test input, respectively.

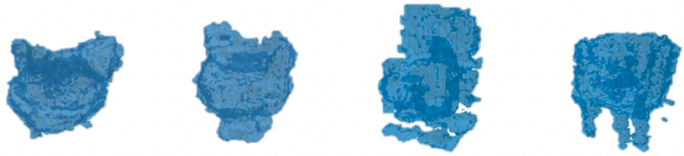
4.3 GAN

GANs are unsupervised learning algorithms that use a supervised loss as part of the training. So their results are expected to be as good as the ones from an AE. Figure 4.3(a) gives some results from a vanilla GAN. We can observe that the vanilla GAN only captures very basic bulky features of chairs but fails on the details, even using a relatively higher resolution. Another disadvantage of the vanilla GAN is that, it can fall into mode collapse easily. In this case, the discriminator is trained too well to classify generated models too easily, thus the generator does not learn anything at all.

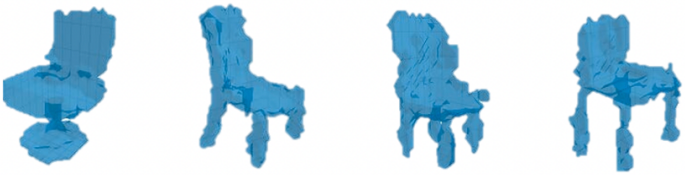
As discussed in Section 3, we are adopting the WGAN method with gradient penalty to overcome the aforementioned problems. Since we are using WGAN,

the last sigmoid layer of the discriminator has been removed. No log operations were used for both losses. Besides gradient penalty, noise was added by doing interpolation between generated and real data before feeding into the discriminator. Figure 4.3(b) gives some optimal results obtained in our experiments. The latent representations we used were sampled from a distribution of $N(0, 0.33)$. As can be observed from the figures, although the generated objects are not extreme smooth, they are already in decent chair-like shapes.

Experiments with other settings have also been carried out. For example, Figure 4.4(a) gives the results of using original sigmoid layer instead of tanh for the last layer of generator. The model may generate lots of floating artifacts shortly after the training begins. In order to reduce the memory consumption for future architecture update, we tried to half the number of feature maps we used between the layers. From Figure 4.4(b) we can observe that obviously the results are



(a) Reconstruction results from GAN



(b) Reconstruction results from WGAN-GP

Figure 4.3: Reconstruction results of chair objects from (a) vanilla GAN and (b) WGAN-GP. The vanilla GAN only captures very basic bulky features of chairs but fails on the details, even using a higher resolution. WGAN-GP can already generate decent chair-like shapes.

not promising anymore. Regarding the initial parameter distribution of the latent representations, experiments have been done with the more often used distribution of $N(0, 1)$, results are shown in Figure 4.4(c). Apparently the generated objects are more noisy.

(a) Using sigmoid layer for generator, instead of tanh

(b) Using half number of feature maps

(c) Using a latent vector distribution of $N(0, 1)$, instead of $N(0, 0.33)$

Figure 4.4: Reconstruction results of chair objects from WGAN-GP with other different settings. (a) For the last layer of the generator, using sigmoid instead of tanh. (b) Using half number of feature maps in the network. (c) The latent representations are sampled from a distribution of $N(0, 1)$, instead of $N(0, 0.33)$.

4.4 AE/VAE-GAN and more

In previous subsections, we have proven that our AE model and GAN model are working. To learn a better universal representation and achieve better synthesis performance, we may combine those two models since the decoder part in AE is exactly the generator part in GAN. However, during the actual testing, this idea never worked if they are straightforwardly combined. The main problem of this idea is that the learned universal representation with AE does not naturally follow a Gaussian distribution, while it is a mandatory requirement for GAN as the input. Hence, Variational Autoencoder (VAE) should be used here for the integration. For the encoder of an Autoencoder, each input is mapped directly to one point in the latent space, which leads to the discontinuous latent space and huge gaps between groups of similar points from the input space. In a variational autoencoder, each input is instead mapped to a multivariate normal distribution around a point in the latent space, which makes a continuous latent space. Continuous latent space also makes the generation of new 3D object possible and the analysis of latent space easier.

On the other hand, this report is mainly about the learning process of universal representations. Reconstructed objects are used to validate the effectiveness of proposed method. In order to make it more illustrative, experiments regarding the investigations in the latent space should be carried out in future. For example, not only for the synthesis tasks, but also for the analytical tasks including object classification.

5 Conclusion and Outlook

In this report, we proposed a switch autoencoder method to learn universal latent representations for 3D object with multiple-formats input. Synthesis experiments have been carried out to validate the effectiveness of the proposed method. With the learned universal representation, decoders can generate 3D objects of better quality. The next step of our future experiments is to make VAE and VAE-GAN work, with which the interpretability of the learned latent representation may be explored. As discussed in Section 4.4, more experiments will be done regarding

the latent space, e.g. similarity search or shape interpolation. In the future, other 3D formats like point cloud may be included. Semantic information may also be used here for better interpretable latent representation learning.

References

[1] Martín Arjovsky, Soumith Chintala, and L. Bottou. "Wasserstein GAN". In: *ArXiv* abs/1701.07875 (2017).

[2] A. Brock et al. "Generative and Discriminative Voxel Modeling with Convolutional Neural Networks". In: *ArXiv* abs/1608.04236 (2016).

[3] C. Choy et al. "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction". In: *European Conference on Computer Vision (ECCV)* (2016).

[4] Y. Feng et al. "GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 264–272.

[5] Sambit Ghadai et al. "Multi-Level 3D CNN for Learning Multi-Scale Spatial Features". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019), pp. 1152–1156.

[6] Rohit Girdhar et al. "Learning a Predictable and Generative Vector Representation for Objects". In: *European Conference on Computer Vision (ECCV)* (2016).

[7] I. Goodfellow, Yoshua Bengio, and Aaron C. Courville. "Deep Learning". In: *Nature* 521 (2015), pp. 436–444.

[8] Ian J. Goodfellow et al. "Generative Adversarial Networks". In: *ArXiv* abs/1406.2661 (2014).

[9] JunYoung Gwak et al. "Weakly Supervised 3D Reconstruction with Adversarial Constraint". In: *2017 International Conference on 3D Vision (3DV)* (2017), pp. 263–272.

[10] Vishakh Hegde and R. Zadeh. "FusionNet: 3D Object Classification Using Multiple Data Representations". In: *ArXiv* abs/1607.05695 (2016).

[11] Diederik P. Kingma and M. Welling. “An Introduction to Variational Autoencoders”. In: *Found. Trends Mach. Learn.* 12 (2019), pp. 307–392.

[12] Jerry Liu, F. Yu, and T. Funkhouser. “Interactive 3D Modeling with a Generative Adversarial Network”. In: *2017 International Conference on 3D Vision (3DV)* (2017), pp. 126–134.

[13] D. Maturana and S. Scherer. “VoxNet: A 3D Convolutional Neural Network for real-time object recognition”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2015), pp. 922–928.

[14] C. R. Qi et al. “Volumetric and Multi-view CNNs for Object Classification on 3D Data”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 5648–5656.

[15] O. Sorkine-Hornung. “Laplacian Mesh Processing”. In: *Eurographics* (2005).

[16] Hang Su et al. “Multi-view Convolutional Neural Networks for 3D Shape Recognition”. In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 945–953.

[17] J. Sun, M. Ovsjanikov, and L. Guibas. “A Concise and Provably Informative MultiScale Signature Based on Heat Diffusion”. In: *Computer Graphics Forum* 28 (2009).

[18] Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. “Multi-view Consistency as Supervisory Signal for Learning Shape and Pose Prediction”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 2897–2905.

[19] H. Wang et al. “Global-to-local generative model for 3D shapes”. In: *ACM Transactions on Graphics (TOG)* 37 (2018), pp. 1–10.

[20] X. Wei et al. “Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect”. In: *ArXiv* abs/1803.01541 (2018).

[21] Jiajun Wu et al. “Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling”. In: *ArXiv* abs/1610.07584 (2016).

[22] Z. Wu et al. “SAGNet”. In: *ACM Transactions on Graphics (TOG)* 38 (2019), pp. 1–14.

[23] Zhirong Wu et al. “3D ShapeNets: A deep representation for volumetric shapes”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1912–1920.

[24] Luisa M. Zintgraf et al. “Visualizing Deep Neural Network Decisions: Prediction Difference Analysis”. In: *ArXiv* abs/1702.04595 (2017).

Karlsruher Schriftenreihe zur Anthropomatik (ISSN 1863-6489)

Band 1 Jürgen Geisler
Leistung des Menschen am Bildschirmarbeitsplatz. 2006
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. 2010
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. 2010
ISBN 978-3-86644-469-0

Band 5 Thomas Usländer
Service-oriented design of environmental information systems. 2010
ISBN 978-3-86644-499-7

Band 6 Giulio Milighetti
Multisensorielle diskret-kontinuierliche Überwachung und Regelung humanoider Roboter. 2010
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. 2011
ISBN 978-3-86644-609-0

Band 8 Eduardo Monari
Dynamische Sensorselektion zur auftragsorientierten Objektverfolgung in Kameranetzwerken. 2011
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. 2011
ISBN 3-86644-760-8

Band 10 Christian Frese
Planung kooperativer Fahrmanöver für kognitive Automobile. 2012
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. 2012
ISBN 978-3-86644-855-1

Band 12 Miriam Schleipen
Adaptivität und Interoperabilität von Manufacturing Execution Systemen (MES). 2013
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. 2013
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. 2013
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. 2014
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. 2014
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. 2015
ISBN 978-3-7315-0320-0

Band 19 Marco Huber
Nonlinear Gaussian Filtering: Theory, Algorithms, and Applications. 2015
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. 2014
ISBN 978-3-7315-0401-6

Band 21 Todor Dimitrov
Permanente Optimierung dynamischer Probleme der Fertigungssteuerung unter Einbeziehung von Benutzerinteraktionen. 2015
ISBN 978-3-7315-0426-9

Band 22 Benjamin Kühn
Interessengetriebene audiovisuelle Szenenexploration. 2016
ISBN 978-3-7315-0457-3

Band 23 Yvonne Fischer
Wissensbasierte probabilistische Modellierung für die Situationsanalyse am Beispiel der maritimen Überwachung. 2016
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. 2016
ISBN 978-3-7315-0519-8

Band 25 Pascal Birnstill
Privacy-Respecting Smart Video Surveillance Based on Usage Control Enforcement. 2016
ISBN 978-3-7315-0538-9

Band 26 Philipp Woock
Umgebungskartenschätzung aus Sidescan-Sonardaten für ein autonomes Unterwasserfahrzeug. 2016
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. 2017
ISBN 978-3-7315-0580-8

Band 28 Erik Ludwig Krempel
Steigerung der Akzeptanz von intelligenter Videoüberwachung in öffentlichen Räumen. 2017
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. 2017
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. 2017
ISBN 978-3-7315-0642-3

Band 32 David Münch
Begriffliche Situationsanalyse aus Videodaten bei unvollständiger und fehlerhafter Information. 2017
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. 2017
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. 2018
ISBN 978-3-7315-0779-6

Band 35 Michael Grinberg
Feature-Based Probabilistic Data Association for Video-Based Multi-Object Tracking. 2018
ISBN 978-3-7315-0781-9

Band 36 Christian Herrmann
Video-to-Video Face Recognition for Low-Quality Surveillance Data. 2018
ISBN 978-3-7315-0799-4

Band 37 Chengchao Qu
Facial Texture Super-Resolution by Fitting 3D Face Models. 2018
ISBN 978-3-7315-0828-1

Band 38 Miriam Ruf
Geometrie und Topologie von Trajektorienoptimierung für vollautomatisches Fahren. 2018
ISBN 978-3-7315-0832-8

Band 39 Angelika Zube
Bewegungsregelung mobiler Manipulatoren für die Mensch-Roboter-Interaktion mittels kartesischer modellprädiktiver Regelung. 2018
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. 2019
ISBN 978-3-7315-0936-3

Band 41 Marco Thomas Gewohn
Ein methodischer Beitrag zur hybriden Regelung der Produktionsqualität in der Fahrzeugmontage. 2019
ISBN 978-3-7315-0893-9

Band 42 Tianyi Guan
Predictive energy-efficient motion trajectory optimization of electric vehicles. 2019
ISBN 978-3-7315-0978-3

Band 43 Jürgen Metzler
Robuste Detektion, Verfolgung und Wiedererkennung von Personen in Videodaten mit niedriger Auflösung. 2020
ISBN 978-3-7315-0968-4

Band 44 Sebastian Bullinger
Image-Based 3D Reconstruction of Dynamic Objects Using Instance-Aware Multibody Structure from Motion. 2020
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. 2020
ISBN 978-3-7315-1028-4

Band 46 Stefan Becker
Dynamic Switching State Systems for Visual Tracking. 2020
ISBN 978-3-7315-1038-3

Band 47 Jennifer Sander
Ansätze zur lokalen Bayes'schen Fusion von Informationsbeiträgen heterogener Quellen. 2021
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. 2021
ISBN 978-3-7315-1082-6

Band 49 Thomas Emter
Integrierte Multi-Sensor-Fusion für die simultane Lokalisierung und Kartenerstellung für mobile Robotersysteme. 2021
ISBN 978-3-7315-1074-1

Band 50 Patrick Dunau
Tracking von Menschen und menschlichen Zuständen. 2021
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. 2021
ISBN 978-3-7315-1091-8

Lehrstuhl für Interaktive Echtzeitsysteme
Karlsruher Institut für Technologie

Fraunhofer-Institut für Optronik, Systemtechnik
und Bildauswertung IOSB Karlsruhe

In 2020, 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 the IOSB in Karlsruhe for the first time due to the pandemic and the strict regulations of such an event in this times. For a week from the 27th to the 31st July the doctoral students of both institutions presented extensive reports on the status of their research and discussed topics ranging from computer vision and optical metrology to network security, usage control and machine learning. 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 978-3-7315-1091-8