Karlsruher Schriften
zur Anthropomatik

Band 45

Jürgen Beyerer, Tim Zander (Eds.)

# Proceedings of the 2019 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

KIT Scientific Publishing

Jürgen Beyerer, Tim Zander (Eds.)

# Proceedings of the 2019 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory

Karlsruher Schriften zur Anthropomatik
Band 45
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 2019 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 2019 – Gedruckt auf FSC-zertifiziertem Papier

ISSN 1863-6489
ISBN 978-3-7315-1028-4
DOI 10.5445/KSP/1000118012

# Preface

In 2019, 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) has again been hosted on the Griesgethof nearby the town of Triberg-Nussbach in Germany.

For a week from July, 29 to August, 2 the PhD students of the both institutions delivered extended reports on the status of their research and participated in thorough 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. Special thanks goes to Prof. Dr. Stephan Klaus from the Mathematical Research Institute of Oberwolfach for giving us a very inspiring tour through the MiMa, Museum for Minerals and Mathematics on the excursion day of the workshop.

The editors thank Julius Krause, Florian Becker, Arno Appenzeller, Paul Wagner and 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

# Towards a Privacy Compliant Research Interface for Multicenter Medical Data

*Arno Appenzeller*

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

Technical Report IES-2019-06

## Abstract

Big Data analysis gains more and more interest in the processing of e-Health data. The potentially big benefit of those analyses comes with a set of new unknown impacts to an individual's privacy. Still it is important to find a balance between privacy impact and utility of the medical data analysis. To achieve this, this technical report takes a look on different privacy preserving techniques, that could be used for a privacy preserving research interface for medical data. The three techniques Differential privacy, $k$-Anonymity and Secure multi-party Computation are evaluated on their feasibility for a medical use-case. With those preliminaries some formal definitions are made for a privacy preserving research interface which implements an hybrid approach of the three techniques and a consent based interface.

## 1 Introduction

The digitization in the health care sector is starting to gain more and more traction. As a consequence of the digitization more e-Health data than ever

before is accessible for broad use cases. As the amount of data to a given topic is growing, Big Data research usually start to become interested in those topics. Especially for medical data Big Data promises new therapies and new valuable insights on different diseases [12]. A more or less open question from a technical perspective is data protection regarding medical data. From the law perspective, for example with the European General Data Protection Regulation (GDPR), there is a firm opinion on privacy of medical data. However there are many open question when processing large amount of medical data. In general the GDPR categorizes personal health information as special data. Article 9 Paragraph 1 says: "*Processing of [...] data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited*" [3]. At first this means that the processing of personal health data is not allowed. But Article 9 Paragraph 1 a) to j) has exclusions, which allow the processing of this special category of data. One of these exclusions is, if the affected person consents to the usage of their data. Other reasons that allow the processing, like the processing for public interest, are more ambiguous than the explicit consent. While the GDPR asks for an explicit permission for the use of the data from an affected person, even the processing of a large amount of anonymized data does not guarantee privacy. Furthermore a recent study showed that the combination of 15 different attributes per dataset is enough to identify an exact person in the US [10]. This proofs that even if data is only processed in an anonymized way, additional measures have to be taken if an affected individual does not explicitly consent to a certain risk of de-identification.

Another fact we face when working with medical data is that the data environments are often multi centric. This means that the data of a single patient is split across different clinicians or hospitals. As a consequence data from multiple sites need to be coordinated, which means in most cases that a trusted party is needed as a broker for the data. Furthermore the privacy of the data is an important questions when coming from different sources and the data is potentially used in different sites for different purposes. Besides the challenge of a research interface for multi centric health data, there are other challenges like how to merge the data of a single patient from different sites or how the different data providers can be connected securely. However for this technical report we focus on a potential research interface for multicenter medical data. A

main requirement for this is the privacy compliant processing of the personal health information. While maintaining this and providing anonymized/pseudonomynized data when needed, another important thing is to provided a back channel for potential results out of the processed data. Especially if they have important results for an individual.
In this technical report we will have an in-depth look at various techniques to provide privacy on personal data in big datasets while still retaining maximum data precision. Afterwards we will present a concept that combines those mechanism with additional techniques that consider consent to provide a privacy preserving research interface for multicenter medical data. In the end this concept will be concluded and an outlook is provided.

# 2 Related work

Like mentioned in the introduction a recent study by Rocher et al. showed that 15 different attributes are enough to identify 99.8% of the citizen of Massachusetts [10]. The claim is proven with a statistical model. This applies regardless how incomplete the data is, so anonymization will not provide enough benefit to protect an individuals privacy. So even a training set for a machine learning algorithm can be a privacy risk. Because of this conclusion the authors demand for even higher measures, than for example the GDPR demands, to protect the privacy of individuals.
The project "PAPAYA: A Platform for Privacy Preserving Data Analytics" focuses more on the specific issue of a privacy preserving research interface for medical data [2]. Ciceri et al. introduce a project to create privacy-preserving neural networks. The approach uses a combination of encryption, secure multi-party computation, differential privacy and functional encryption. Different data sources are used to train a neural network. The training data is discarded afterwards. All in all they do not provide an in-depth look of their approach. But they present the idea of using differential privacy for the training data to add noise to the original data.
Another project that provides a research interface for medical data is the MOSAIC project [1]. Bialke et al. describe this in "*MOSAIC - A Modular Approach to*

*Data Management in Epidemiological Studies*". The authors want to comply with privacy requirements by using study-specific pseudonymisation and giving access for third parties only through a designated interface. An interesting fact about MOSAIC is that it enables a designated back channel for research algorithm. With this the algorithm can give back individual findings that occurred during the processing. Unfortunately the concept is not explained in more detail.

# 3 Privacy preserving techniques for multicenter environments

The following section presents three techniques that can preserve privacy for large databases. Therefore they can be used for multicenter environments. Finally the three techniques will be evaluated by criteria like accuracy and privacy guarantees.

## 3.1 Differential privacy

In 2006 Dwork et al. introduced the notion of $\varepsilon$-Differential Privacy [6]. In general Differential Privacy has the goal for a certain data in a statistical database to achieve the same level of privacy as if the data is removed or never was in the database. This means that the data of a single individual needs to be modified so that the individual can not be identified. With this approach privacy can be preserved while still retaining a good utility for the processing of the modified data. The assumption for Differential Privacy is, that the likelihood that there is any disclosure, is a very small number regardless if the data is in the database or not. To be more specific the $\epsilon$ in $\epsilon$-Differential Privacy describes the privacy loss when a dataset is released from a database. Therefore a really small $\epsilon$ is desired but certainly it remains important to keep the utility of the data. Formally $K$ is a $\epsilon$-Differential Privacy algorithm if the following is valid: All available data are part of the set $S$. $D_1$ and $D_2$ are datasets that have the difference of at most one element.

**Definition 3.1.1** ($\varepsilon$-Differential Privacy Algorithm)**.**

$$Pr[K(D_1) \in S] \leq e^{\epsilon} * Pr[K(D_2) \in S] \tag{3.1}$$

The conclusion of this definition is that even if data is removed no output and its consequences in regard of privacy loss becomes significantly less or more likely. Which ultimately means that it does not matter if data is in or not in a database, if $K$ fulfills the requirement of Definition 3.1.1.
With this strong privacy guarantees can be achieved but an important factor is the size of the dataset: The smaller the database the higher the noise added (or the smaller $\epsilon$) has to be to alter/randomise the original data.
Another important question is what a good Differential Privacy algorithm is. This question can not be answered in general because it depends on the use case. If the use case is to process numeric values for statistical operations like sum, median or average a good choice is Laplacian noise. This uses the Laplacian mechanism to add noise to the input data. For this algorithm the $\epsilon$ is a measure for the randomization. If $\epsilon = 0$ the privatized data is complete random noise. While in theory this provides obviously the best privacy, the data has no more real utility and leads the Differential Privacy approach ad absurdum.
Differential Privacy can be divided in two different variants. The one is Global Differential Privacy where all original data is stored globally. Only the output of this original data is aggregated to fulfill the requirements of Differential Privacy. For this approach a trusted third party which manages the data is essential. The other variant is Local Differential Privacy. Here every individual or data owner modifies the data before it leaves the origin, so that the original information is nowhere else. For this no trusted third party is needed because the data is already modified when it reaches another party. Besides $\epsilon$-Differential Privacy there also exists ($\epsilon$,$\delta$)-Differential Privacy. This version of Differential Privacy accepts deviations by $\delta$ from the original notion like in Definition 3.1.1.
Differential Privacy is a concept that sounds very promising in theory. While there are practical use cases (even Apple [5] and Google [8] are using it in their mobile systems) the real utility depends on the scenario it is used. There is a review paper by Dankar et al. which provides an in-depth look at medical applications but still the conclusion is that besides statistical evaluations it is

very limited [4]. However for a combination of different techniques Differential Privacy is one of the most promising ones.

## 3.2 $k$-Anonymity

Another technique to preserve privacy is $k$-Anonymity. The method was introduced in 2002 by Sweeny et al. [11]. The main principle of $k$-Anonymity is to alter the existing data of a database, so that they still have utility but it is guaranteed that affected individuals with data in the database can not be reidentified. A collection of datasets can be called $k$-anonymous if one of the datasets ca not be distinguished from $k-1$ other datasets.

**Example 3.2.1** (4-Anonymity)**.** A $k = 4$ anonymized dataset has at least 4 records for each value combination of certain attributes that $k$-Anonymity applies to.

There are two methods to achieve $k$-Anonymity:

- **Suppression**: Parts of the data will be removed, disguised or made indistinguishable (Mapping all data to the same pseudonym e.g.).
- **Generalization**: Modify parts of the data to ranges of values instead of exact values or assign attributes to a more general type.

One issue with $k$-Anonymity is that there is no general measurement for the privacy guarantee. Furthermore additional domain knowledge is required for suppression or generalization of the data. In some cases there are guidelines that could be used for generalization. For example the Canadian Institute for Health Research published the "*CIHR Best Practices for Protecting Privacy in Health Research*" which helps to generalize medical data.
A medical use case for $k$-Anonymity is described by El Emam et al. [7]. Here the previous mentioned guidelines from the Canadian Institute for Health Research are used as background knowledge for an algorithm that generalizes medical data. With this the generalization can be performed automatically and it is also possible to measure the information loss compared to the original data. So

the privacy impact on a dataset to which the guidelines apply can be reduced. They also show real world feasibility of the approach by using it to hand over $k$-anonymous data from pharmacies to commercial data brokers. However the the issue of a universal generalization remains and every use case has to be considered individually.

## 3.3 Secure multi-party computation

The main principles of Secure multi-party Computation (SCM) were already introduced by Yao in the 1980s [13]. The basic idea of this was to evaluate data from different parties without revealing the data.
According to Lindell and Pinkas there can be two models to achieve this [9]. In one case there is a trusted third party that evaluates the data for the participating parties. The other case has no third party one can trust with its data. In this case a direct communication between the data is needed and it needs to be ensured that the data already leaves the participating parties in a private state.
The typical scenario for SMC is that there are several parties that own private data. All parties want to evaluate their data to a common public result. This can also mean that a third party like a research institute gets this data to do the evaluation. The main issue in this scenario is that there is no trust established between the parties or the parties do not want to reveal their data. A special variation of this scenario exists when there is a third party that does the data processing and returns the value to the parties. However for a medical use case it still remains important that the participating parties do not get the raw data but only the final result.
A concrete example for such a scenario is to calculate the average salary of three parties. When using the secret sharing the typical procedure is that the starting party chooses a secret $r$. This secret is added to the own salary $x$ and the result will be sent to the second party. The second party adds its salary $y$ and sends it to party number three which follows the same procedure. This can be easily extended to an arbitrary number of parties. Finally after the round trip the first party gets the result back and subtracts $r$ to receive the final value to calculate the average without revealing its salary to the others or gaining knowledge of the others salary.

Another approach to this is using homomorphic encryption. In this case certain mathematical operations can be done with the ciphertext without knowing the secret key or the need to decrypt it. The operations depend on the homomorphic properties of the encryption method. For example an additive homomorphic property would mean that it is possible to calculate $Enc(a) + Enc(b) = Enc(a + b)$. It needs to be considered that for plain encryption those methods would have a lot of weaknesses to adversaries, but the measures are enough to preserve data privacy. A possible scenario for this would be a third party research algorithm that does a cohort analysis for a clinician. For this it needs the data from the clinician and other participants that provide the comparison data to create the cohort. A main requirement is that the third party does not see the plain data. To realize this a key broker is required which gives a common key to all participants. With the resulting chiphertexts the third party algorithm can do its cohort analysis using the homomorphic properties.
An obvious advantage to the previous techniques is correctness of the result which also implies precision. That means while the results achieved with Differential Privacy or $k$-Anonymity can differ to a certain degree from the real result, SMC always returns the exact result. An issue with SMC is that it has a big overhead in terms of run time. Even simple operations can use a lot of time.

## 3.4 Evaluation of the techniques

After the introduction of the three different techniques considered in this report, we will do an evaluation of them that considers the strengths and weaknesses of the techniques. Table 3.1 gives an overview of this.
In terms of privacy guarantees both Differential Privacy and $k$-Anonymity have metrics that make a statement about the degree of privacy. SMC's guarantees are dependent on the encryption mechanism used and can not be generalized. Full accuracy is provided when using SMC while the privacy preserving mechanism does not rely on modification of the data. Differential Privacy's accuracy is affected by the choice of $\epsilon$, where a very large $\epsilon$ provides good accuracy but not much privacy. For $k$-Anonymity no general assumption can be made because the accuracy depends on the generalisation/surpression method. When considering scalable performance Differential Privacy as well as $k$-Anonymity

should provide good results regardless the amount of data while SMC has a lot of overhead because of the encryption mechanism. Lastly it is important if any kind of trusted party is needed to perform the techniques. Differential Privacy and $k$-Anonymity require a party the manages the data. If considering Differential Privacy it is possible that the local approach is used so the trusted party is only needed for the global approach. Only SMC offers the option to operate completely without a trusted party, if the participants communicate directly with their ciphertext.

**Table 3.1**: Overview of privacy preserving techniques

| | *Techniques* | | |
|---|---|---|---|
| | **Differential Privacy** | **Secure multi-party Computation** | **$k$-Anonymity** |
| *Privacy guarantuees* | • | | • |
| *High Accuracy* | ○ | • | |
| *Scaleable performance* | • | | • |
| *Trusted Party needed* | Partly | No | Yes |
| *Limitations* | Choice of $\epsilon$ affects properties | Utility and processing time heavily depends on the type of SMC | Requires domain knowledge |

# 4 A privacy compliant research interface

To define a research interface it is important to understand the difference between a non-interactive interface and an interactive one. A non-interactive research interface is one where the data is released once and for all and there is no way to modify the data for a certain request. An interactive research interface can decide the privacy strategy for each query since only the data for the given request is released and the complete data remains hidden through the interface.

We think that for a privacy preserving research interface it is important not to follow an one fits all approach. There are different kind of queries that can require different degrees of accuracy. The main goal should always be: *Preserve as much privacy as possible and lose as less accuracy as possible*. This can only be achieved with a hybrid approach. On the one hand a combination of the previously introduced techniques, that are used for the range of queries where the individual technique, can be used best. On the other hand those techniques all fall in specific use cases and can reach their limit, where no more useful query is possible. Furthermore there can be some requests where both the researcher and the affected person can benefit from data that is not anonymized. You can think of queries that can provide feedback on the individual person. For those queries the person's consent is mandatory.

To include this in the desired fully automated research interface a mechanism is required to map the consent in a digital format. Furthermore this consent should be dynamic so that an affected person can authorize or revoke it at any time. In addition to enable automatic evaluation of this, an enforcement mechanism is needed to evaluate consent for each query. Medical consent in a digital format is a non trivial task with some existing concepts but most of them are far from complete. We will postpone this part which we call *consent based interface* to future work.
We assume that he research interface exposes a set of privacy functions like $\mathbf{P\_SUM}$, $\mathbf{P\_AVERAGE}$, $\mathbf{P\_MEDIAN}$ etc. to do operations on attributes of the data in the database.

**Definition 4.0.1** (Privacy preserving functions)**.** A privacy preserving research interface defines a Set $\mathcal{F}$ of privacy preserving function. They all follow the following naming convention $\mathbf{P\_}*$ where $*$ is a mathematical function like $\mathbf{SUM}$ or $\mathbf{COUNT}$.

To perform a query the researcher has to provide additional properties. It needs to be defined if $accuracy$ or $privacy$ to which scale is desired or if an algorithm wishes to provide additional feedback to an individual $feedback$.

**Definition 4.0.2** (Privacy preserving configuration)**.** A privacy preserving research interface has a Set $\mathcal{C} = \{accuracy, privacy(x), feedback\}$ which con-

tains the privacy preserving configuration for a request. $privacy(x)$ has $x \in \mathbb{N}^+$ as number to indicate the factor of the privacy impact.

With this a request can be formulated. Such a request uses a query language in an interface specific language where the request attributes from the dataset can be defined. A privacy function out of $\mathcal{F}$ also needs to be used in this query. In addition a configuration needs to be provided to indicate what the requirements for the request are.

**Definition 4.0.3** (Privacy preserving request)**.** A request $req$ for a privacy preserving research interface looks like the following: $req = (query, config)$ where $query$ is a query made with a query language $\mathbf{QL}$ that includes $\mathcal{F}$ and $config \in \mathcal{C}$.

With such a request $req$ the interface can now decide depending on $config$ which privacy preserving technique should be used. The following Definition 4.0.4 illustrates this.

**Definition 4.0.4** (Evaluation of $config$)**.**

$$config = \begin{cases} \text{if } accuracy \rightarrow \text{use SMC} \\ \text{if } privacy(x) \rightarrow \text{use Differential Privacy} \\ \hookrightarrow \text{or } k\text{-Anonymity depending on } x \\ \text{if } feedback \rightarrow \text{use } \textit{consent based interface} \end{cases}$$

# 5 Conclusion & outlook

This technical report looks at three different techniques to preserve privacy on an individuals data. All of these three techniques have various advantages and disadvantages. While Differential Privacy and $k$-Anonymity have good privacy guarantees they can lack accuracy. SMC can provide accuracy on the results but its performance can be a great uncertainty. So there is certainly no one fits all approach. In fact a hybrid approach that combines those three techniques and that chooses the best depending on the requirements for a certain request is proposed. In addition there can be requests where those techniques can not help

or do not fit the requirement. Therefore a fallback to the individuals consent is needed. With this definition of a privacy preserving research interface for multicenter medical data the foundation for more in-depth work and experiments with real world e-Health data is made.

While this report provides the fundamentals a real world evaluation needs to be done. It needs be proven that the introduced privacy preserving techniques work good on real medical data. Another issue that remains is a good privacy metric. This is especially required for an informed consent decision of a patient. Considering that the consent based interface needs to be introduced in future work. With this integration a full feature research interface is possible, which remains open for further refinement. Finally this approach should be evaluated against the GDPR. It has to be figured out what is needed to be compliant to it and what an interface should provide to fulfill requirements of the GDPR.

# References

[1] M. Bialke et al. “MOSAIC – A Modular Approach to Data Management in Epidemiological Studies”. In: *Methods of Information in Medicine* 54.04 (2015), pp. 364–371.

[2] Eleonora Ciceri. “PAPAYA: A Platform for Privacy Preserving Data Analytics”. In: *ERCIM News* 118 (2019).

[3] European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. URL: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[4] Fida K. Dankar and Khaled El Emam. “Practicing Differential Privacy in Health Care: A Review”. In: *Trans. Data Privacy* 6.1 (2013), pp. 35–67.

[5] Apple Differential Privacy Team. “Learning with Privacy at Scale”. In: *Apple Machine Learning Journal* 1.8 (2017).

[6] Cynthia Dwork. “Differential Privacy”. In: (2006), pp. 1–12.

[7] K El Emam et al. “A globally optimal k-anonymity method for the de-identification of health data.” In: *J Am Med Inform Assoc* 16.5 (2009), pp. 670–682.

[8] M. Guevara. *Enabling developers and organizations to use differential privacy*. 2019. URL: https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html (visited on 10/25/2019).

[9] Yehuda Lindell and Benny Pinkas. “Privacy Preserving Data Mining”. In: *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology*. CRYPTO ’00. London, UK, UK: Springer-Verlag, 2000, pp. 36–54. ISBN: 3-540-67907-3. URL: http://dl.acm.org/citation.cfm?id=646765.704129.

[10] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. “Estimating the success of re-identifications in incomplete datasets using generative models”. In: *Nature Communications* 10.1 (2019).

[11] Latanya Sweeney. “K-anonymity: A Model for Protecting Privacy”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. URL: http://dx.doi.org/10.1142/S0218488502001648.

[12] K. Verspoor and F. Martin-Sanchez. “Big Data in Medicine Is Driving Big Changes”. In: *Yearbook of Medical Informatics* 23.01 (2014), pp. 14–20.

[13] A. C. Yao. “Protocols for secure computations”. In: *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 1982, pp. 160–164.

# Semi-Supervised Manifold Learning for Hyperspectral Data

*Florian Becker*

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

Technical Report IES-2019-11

## Abstract

There are real world data sets where a linear approximation like the principal components might not capture the intrinsic characteristics of the data. Nonlinear dimensionality reduction or *manifold* learning uses a graph-based approach to model the local structure of the data. Manifold learning algorithms assume that the data resides on a low-dimensional manifold that is embedded in a higher-dimensional space. For real world data sets this assumption might not be evident. However, using manifold learning for a classification task can reveal a better performance than using a corresponding procedure that uses the principal components of the data. We show that this is the case for our hyperspectral data set using the two manifold learning algorithms Laplacian eigenmaps and locally linear embedding.

## 1 Introduction

Nonlinear dimensionality reduction or *manifold learning* is a useful tool for high-dimensional data analysis. In contrast to linear dimensionality reduction

as it is performed by a standard principal component analysis (PCA), with manifold learning the possibly low-dimensional manifold that is embedded in a high-dimensional space can be uncovered. This so-called *manifold assumption* is central to the theory of manifold learning and states that the data resides on a low-dimensional manifold in high-dimensional space. Manifold learning has been applied to many computer vision problem domains including face recognition [3], image retrieval [4] and medical image analysis [1]. Due to the high spectral resolution of many hyperspectral image data sets and the high correlation between adjacent and overtone bands, manifold learning has received some attention in the research community [5].
In this technical report, we will first review the basics of manifold learning, why it is a useful framework and how it can be utilized for classification in a semi-supervised manner. Finally, we will apply this semi-supervised procedure to a hyperspectral data set consisting of four different kinds of wood (chips): eucalyptus, poplar, beech and spruce. The results indicate that manifold learning outperforms a linear approach using PCA.

# 2 Classification with manifold learning

Discovering the low-dimensional manifold embedded in a higher-dimensional space can be utilized for classification. We aim to show two aspects of manifold learning: First, it can be employed for classification, second, manifold learning outperforms a corresponding linear procedure using principal component analysis. In general, dimensionality reduction is often used as a step prior to classification. This is due to the fact that for many datasets, the dimensions of individual data points might be correlated due to the physical nature of the process that has generated the data. For instance, in (near) infrared spectroscopy overtone bands can be observed that are a manifestation of the vibrational modes. As the resonant frequencies can be approximated by an harmonic oscillator, characteristic peaks in the spectrum might arise from the vibrational modes of the same chemical substance. For a classification task correlation means that specific dimensions might not carry valuable information, in the sense that the additional information does not lead to a better separability of the data and

therefore also does not contribute to the classification performance. Removing correlated dimensions can therefore lead to a simpler classifier with less parameters. When applying manifold learning prior to classification, the objective is to exploit the manifold assumption. Manifold learning is a good fit to the data when there are non-linear dependencies between different dimensions. In practice, it is not evident that non-linear dependencies exist in high-dimensional data. However, if manifold learning leads to better classification results than a linear method, this might indicate the presence of an intrinsic low-dimensional manifold.

# 3 Laplacian eigenmaps

We briefly review the basics of one popular manifold learning algorithm called Laplacian Eigenmaps (LE). Given data samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=0}^{N} \subseteq \mathbb{R}^n$, LE computes a Laplacian matrix according to a kernel function. The final mapping is then defined by the eigenvectors of the graph Laplacian matrix. A detailed description is given by Algorithm 3.1 below. Central to the algorithm is the choice of the kernel function. We call a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel, if the induced *Gram matrix* defined by $K_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$ is positive semi-definite, i.e.

$$\boldsymbol{x}^T \boldsymbol{K} \boldsymbol{x} = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i \, K_{ij} \, x_j \geq 0, \tag{3.1}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$. This is the discrete analog to Mercer's condition [6] which states that the function $K : [a, b] \times [a, b] \to \mathbb{R}$ fulfills the inequality

$$\int \int f(x) K(x, y) f(y) dx dy \geq 0 \tag{3.2}$$

for every function $f \in L^2(\mathbb{R})$. A symmetric kernel function satisfying Mercer's condition leads to nonnegative real eigenvalues and orthogonal eigenvectors for the corresponding kernel matrix.

**Algorithm 3.1** Laplacian Eigenmaps [2]

**procedure** LAPLACIAN EIGENMAPS
  **Input:** data $\mathcal{X} = \{\mathbf{x}_i\}_{i=0}^{N} \subseteq \mathbb{R}^n$
  **Output:** embedding $\mathcal{Y} = \{\mathbf{y}_i\}_{i=0}^{N} \subseteq \mathbb{R}^m$
  1.) Build an adjacency graph $G = (V, E)$
    nodes $v_i \in V$ and $v_j \in V$ are connected if $||\boldsymbol{x_i} - \boldsymbol{x_j}||_2^2 < \varepsilon$
  2.) Pick weights
    Choose a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and set

$$\boldsymbol{W}_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j) & (i,j) \in E \\ 0 & \text{else} \end{cases}$$

  3.) Compute Eigenmap
    $\boldsymbol{L}\boldsymbol{y} = \lambda \boldsymbol{D}\boldsymbol{y}, \quad$ with $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{W}_{ji}$ and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$
    $\boldsymbol{x}_i \to (\boldsymbol{y}_1(i), \ldots, \boldsymbol{y}_m(i))$
**end procedure**

The embedding is found by computing the generalized eigenvalue problem involving the graph Laplacian and the corresponding degree matrix. The nonlinear nature of Algorithm 3.1 is due to the choice of the kernel function.

# 4 Semi-supervised manifold learning

Semi-supervised machine learning methods make use of unlabeled data points for training. Transductive learning is one variant of a semi-supervised learning setting where the correct labels of some given unlabeled data points must be inferred. This is in contrast to inductive learning where a function is learned that maps a data point to its label. Manifold learning algorithms are label-agnostic: In order to build the adjacency graph no information about class labels is necessary. The main idea behind a semi-supervised manifold learning approach is that the kernel matrix is built using labeled and unlabeled data points. The resulting matrix quantifies the similarity between all pairwise data points. As a subset of these data points is labeled, the kernel matrix relates each unlabeled data point to every labeled data point. The computation of the eigenmap and

the projection of the high-dimensional data leads to an embedded space with partially labeled data points. Unlabeled points can be classified with a simple nearest-neighbor search. In this way, the intrinsic manifold structure–given that it exists–is put to use for a classification task.

**Algorithm 4.1** Semi-Supervised Manifold Learning

**procedure** SEMI-SUPERVISED MANIFOLD LEARNING
  **Input:** labeled data $\{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_p, c_p)\}$,
    unlabeled data $\{\boldsymbol{x}_1^u, \ldots, \boldsymbol{x}_q^u\}$
  **Output:** labels for $\{\boldsymbol{x}_1^u, \ldots, \boldsymbol{x}_q^u\}$
  1.) Compute embedding by manifold learning algorithm
    e.g. by $\boldsymbol{L}\boldsymbol{y} = \lambda \boldsymbol{D}\boldsymbol{y}$
  2.) Embed all data points
    $\boldsymbol{x}_i \longrightarrow \boldsymbol{y}_1(i), \ldots, \boldsymbol{y}_m(i)$
  3.) Classify unlabeled data points
  **for all** unlabeled data points $\boldsymbol{x}^u$ **do**
    get the labels of the $k$ nearest labeled points in the embedded space
    assign data point $\boldsymbol{x}^u$ the most common label
  **end for**
**end procedure**

The procedure described above can be used together with any manifold learning algorithm. In order to compare LE, we also apply a further manifold learning algorithm to the data set called locally linear embedding (LLE). For a given data set $\mathcal{X} = \{\mathbf{x}_i\}_{i=0}^N \subseteq \mathbb{R}^n$, LLE tries to reconstruct every data point from a linear combination of its $k$-nearest neighbors. LLE minimizes the following cost function:

$$\begin{aligned} E(W) &= \sum_{i=1}^{N} ||\boldsymbol{x}_i - \sum_{\boldsymbol{x}_j \in \mathcal{N}_k(\boldsymbol{x}_i), j \neq i} w_{ij} \boldsymbol{x}_j||_2^2 \\ \text{s.t.} \sum_{i=1} W_{ij} &= 1 \qquad \forall j \in \{1, \ldots, N\} \end{aligned} \tag{4.1}$$

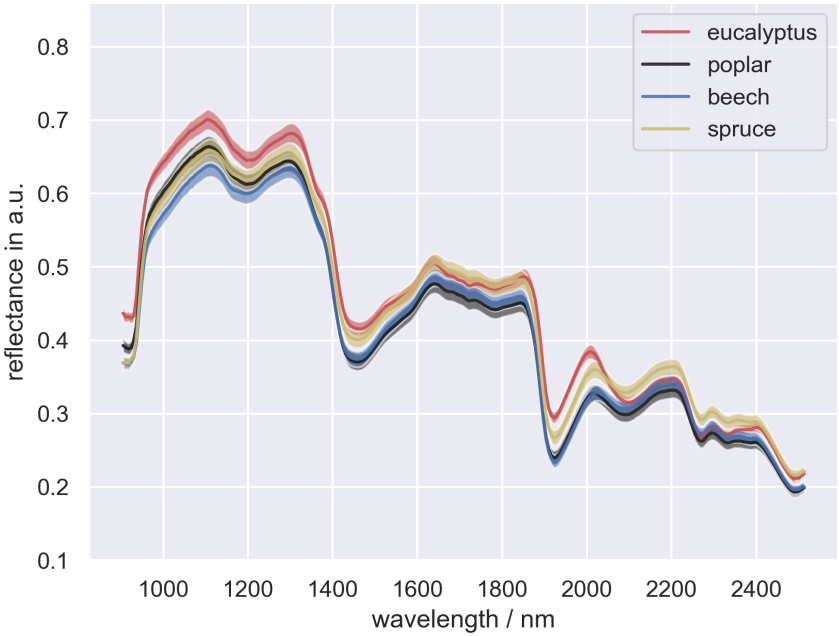



**Figure 4.1**: Spectra of the four different woods: eucalyptus, poplar, beech and spruce. This plot also shows the standard deviation ($0.1\sigma$) around the mean.

$\mathcal{N}_k(\boldsymbol{x})$ denotes the set of $k$-nearest neighbors of $\boldsymbol{x}$. In order to achieve a neighborhood preserving map, the resulting weight matrix from the optimization problem 4.1 above is used to find an embedding:

$$E(Y) = \sum_{i=1}^{N} ||y_i - \sum_{y_j \in \mathcal{N}(y_i), j \neq i} w_{ij} y_j||_2^2. \tag{4.2}$$

In the following, we describe the methodology that was used to apply and validate Algorithm 4.1 for hyperspectral data. The hyperspectral images were acquired using a Specim SWIR camera with spectral range from 950 nm–2500 nm and a spectral resolution of 10 nm. Figure 4.1 shows the entirety of the

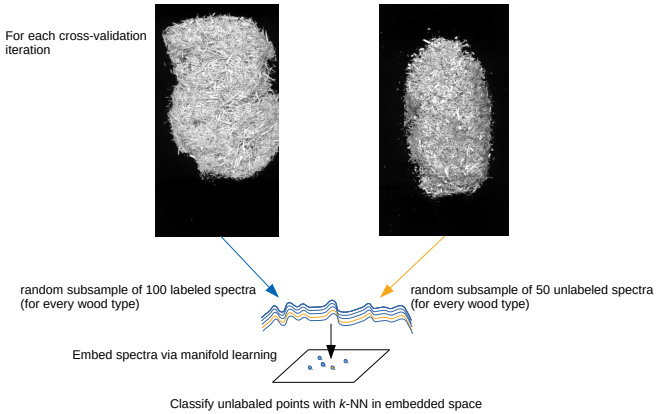



**Figure 4.2**: The proposed methodology. We acquired separate data sets for training and testing. For each cross-validation iteration, we sampled 100 labeled and 50 unlabeled spectra from every wood type. No further preprocessing of the spectra is applied. Based on this data, the Laplacian (and the locally linear embedding optimization problem) is computed. The images above of the fine wood chips are averages over all hyperspectral bands.

spectra for the four classes in terms of a mean spectrum with $0.1\sigma$. Separate image sets were acquired for training and testing. To evaluate Algorithm 4.1, a target dimension of 2 was chosen for all dimensionality reduction procedures.

# 5 Results

The above methodology leads to the results given in Table 5.1. The results indicate that the used manifold learning algorithms outperform linear dimensionality reduction in terms of a 1-nearest neighbor classification in the embedded space. Furthermore, we used two different kernel functions $k_{\mathrm{rbf}}$ and $k_{\cos}$. The overall accuracy for $k_{\cos}$ leads to better results. As the spectra were not preprocessed, this result is not too surprising as the $\cos$-similarity is invariant to linear shifts of the spectrum–which is in contrast to the $\mathrm{rbf}$-kernel. We furthermore observe that LE outperforms LLE for our data set.

**Table 5.1**: Classification results for PCA and the different manifold learning algorithms using the the semi-supervised manifold learning procedure outlined in Algorithm 4.1. The overall accuracy (OA) is given in the last column.

| Method | **Eucalyptus** | **Poplar** | **Beech** | **Spruce** | **OA** ($\mu + \sigma$) |
|---|---|---|---|---|---|
| PCA | 0.61 | 0.74 | 0.76 | 0.58 | 0.67 + 0.036 |
| $\text{LLE}_{k=30}$ | 0.68 | 0.81 | 0.76 | 0.60 | 0.71 + 0.019 |
| $\text{LLE}_{k=40}$ | 0.72 | 0.83 | 0.76 | 0.63 | 0.74 + 0.018 |
| $\text{LE}_{\text{rbf}}$ | 0.75 | 0.86 | 0.76 | 0.61 | 0.74 + 0.020 |
| $\text{LE}_{\cos}$ | 0.76 | 0.95 | 0.75 | 0.66 | 0.78 + 0.016 |

Especially $\text{LE}_{\cos}$ significantly outperforms the PCA-based approach. In addition, as indicated by the standard deviation, $\text{LE}_{\cos}$ is the most robust method, while throughout the cross-validation the variance of the PCA-based procedure is the highest.

# 6 Conclusion & Outlook

In essence, Laplacian eigenmaps and locally linear embedding build a discrete approximation of the underlying data manifold. By computing a weight matrix that captures the local structure of the data, the intrinsic characteristics are utilized for dimensionality reduction. The induced neighborhood preserving map is a suitable tool for high-dimensional data analysis. We have applied manifold learning for a semi-supervised classification task and showed that it outperforms classification in the space that is defined by the principal components. Our results indicate that choosing a kernel function is a critical step for LE. Manifold learning has the potential to uncover the low-dimensional manifold of the data. Future work should continue to examine this potential.

# References

[1] Paul Aljabar et al. “Combining morphological information in a manifold learning framework: application to neonatal MRI”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 1–8.

[2] Mikhail Belkin and Partha Niyogi. “Laplacian eigenmaps for dimensionality reduction and data representation”. In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[3] Xiaofei He et al. “Face recognition using laplacianfaces”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 3 (2005), pp. 328–340.

[4] Yen-Yu Lin, Tyng-Luh Liu, and Hwann-Tzong Chen. “Semantic manifold learning for image retrieval”. In: *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM. 2005, pp. 249–258.

[5] Dalton Lunga et al. “Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning”. In: *IEEE Signal Processing Magazine* 31.1 (2013), pp. 55–66.

[6] Alex J Smola and Bernhard Schölkopf. “From regularization operators to support vector kernels”. In: *Advances in Neural information processing systems*. 1998, pp. 343–349.

# An Overview of Ellipsometric Measurements for Nonplanar Surfaces

*Chia-Wei Chen*

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

Technical Report IES-2019-09

## Abstract

Ellipsometry is an optical method used for characterizing materials and thin films. The principle of ellipsometry is that it measures polarization changes at a sample in a reflection or transmission configuration. However, the shape of the sample is limited to flat or nearly flat surfaces because ellipsometry is sensitive to the angle of incidence, tilt angle and the sample position (height). Even slight misalignment of the sample might lead to significant experimental errors. For large misalignment, the detector of the ellipsometer is not feasible to receive sufficient signals. There have been a few approaches for characterizing nonplanr surfaces by ellipsometry. This report gives an overview of these approaches for ellipsometric measurements of nonplanar surfaces.

## 1 Introduction

Ellipsometry is an optical technique for characterization of materials and thin films. The main features of ellipsometry are high precision (thickness from

a few Å to several tens of microns), nondestructive measurement, and wide applications. The principle of ellipsometry is that it measures polarization changes at a sample in a reflection or transmission configuration. Fig. 1.1 shows the principle of reflection ellipsometry. The incident light is linearly polarized. After the reflection from the substrate, the reflected light becomes elliptically polarized. The Fresnel equations describe the interaction of light (electromagnetic waves) and materials. The polarization changes can be defined as the ratio $\rho$ of the amplitude reflection coefficients for p- and s- polarizations [1]:

$$\rho = \frac{r_p}{r_s} = \tan\Psi e^{i\Delta}, \tag{1.1}$$

where $\Psi$ and $\Delta$ present amplitude ratio and phase difference. Ellipsometry technique can be applied to many scientific and industry fields, e.g., semiconductor, chemistry, display industry and biomaterials [7].

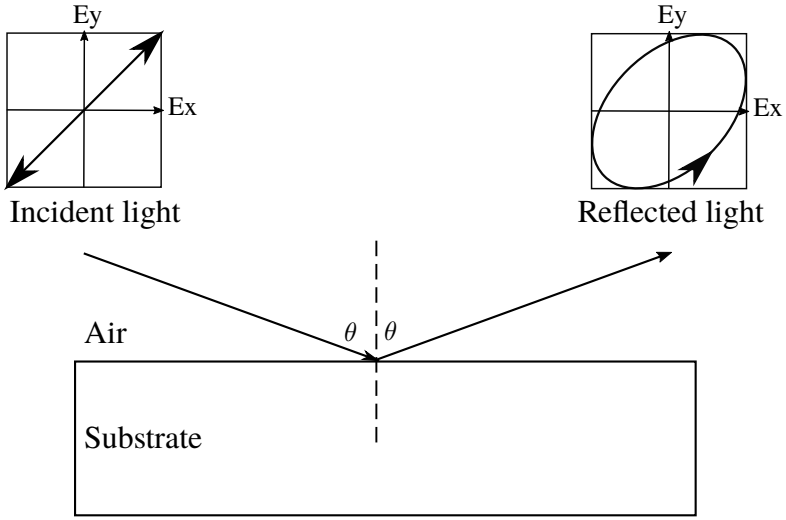



**Figure 1.1**: Measurement principle of ellipsometry.

In conventional ellipsometers, samples are limited to a planar shape because ellipsometry is sensitive to the angle of incidence (AOI), tilt angle and the sample

position (height). Even slight misalignment of the position and orientation of the sample might lead to significant experimental errors. For large misalignment, the detector of the ellipsometer is not feasible to receive sufficient signals. For nonplanar surfaces, the beam path of reflected or transmitted light is changed because of the surface shape. In order to solve this problem, different methods have been proposed for nonplanar surfaces. In this paper, we will review and compare these approaches in the configuration of reflection ellipsometry.

# 2 Surface orientation in reflection ellipsometry

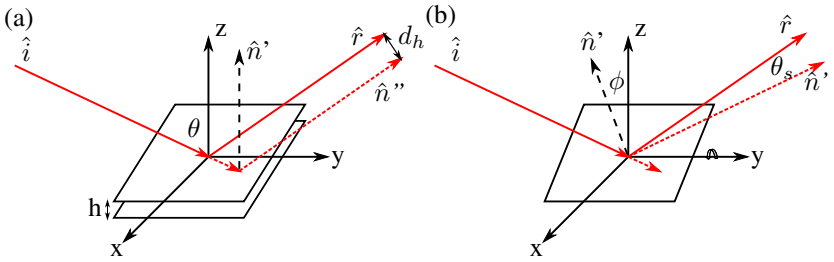



**Figure 2.1**: Definition of the surface orientation. (a) An offset $h$ along the surface normal $\hat{n}$. (b) The surface rotates around the y-axis.

Fig. 2.1(a) shows a planar surface defining the xy-plane. The z-axis $(0, 0, 1)$ is the surface normal $\hat{n}$. The hat is denote as a unit vector. If the incident beam $\hat{i}$ is on the yz plane and the incident angle is $\theta$, the incident beam is expressed as: $(0, \sin\theta, -\cos\theta)$. The reflected beam $\hat{r}$ can be defined as: $(0, \sin\theta, \cos\theta)$. The relationship between $\hat{n}$, $\hat{i}$, and $\hat{r}$ is shown as [11]:

$$\hat{r} = \hat{i} - 2(\hat{i} \cdot \hat{n})\hat{n}. \tag{2.1}$$

The angle of incidence $\theta$ is determined by the surface normal $\hat{n}$ and the incident beam $\hat{i}$ as:

$$\theta = \cos^{-1} -\hat{i} \cdot \hat{n}. \tag{2.2}$$

If the surface has an offset $h$ along the surface normal $\hat{n}$ as shown in Fig. 2.1(a), it will cause an offset $d_h$ of the reflected beam as:

$$d_h = 2h \sin\theta \tag{2.3}$$

For an incident angle of $70^\circ$ and an offset of 1 mm, the offset $d_h$ is about 1.88 mm.

Fig. 2.1(b) illustrates a surface rotates around the y-axis. The surface normal of the sample becomes $\hat{n}' = (\sin\phi, 0, \cos\phi)$. Using Eq. 2.2, we can easily compute the angle of incidence $\theta'$ and the reflected beam $\hat{r}'$ after the rotation as:

$$\cos\theta' = \cos\theta\cos\phi, \tag{2.4}$$

$$\hat{r}' = (\cos\theta\sin 2\phi, \sin\theta, \cos\theta\cos 2\phi). \tag{2.5}$$

The included angle $\theta_s$ between the original reflected beam $\hat{r}$ and the reflected beam $\hat{r}'$ after tilting can be calculated by the product rule from:

$$\cos\theta_s = \sin^2\theta + \cos^2\theta\cos 2\phi. \tag{2.6}$$

For an incident angle of $70^\circ$, if a surface tilts $5^\circ$ around y-axis, it will produce an angle deviation by $3.4^\circ$ for the detector. If the distance between the surface and the detector is 200 mm, it will induce an offset of 11.9 mm.

From the above calculation results, surface offset and tilt produce a significant offset for the detector, which will degrade the measurement accuracy. Therefore, special optical designs, compensation methods and precision alignment are necessary for ellipsometric measurements of nonplanar surfaces.

# 3 Ellipsometric measurements for nonplanar surfaces

There have been a few approaches for characterizing nonplanar surfaces by ellipsometry. These approaches can be categorized into three types: combination of topometry and ellipsometry, polarization model for azimuth deviations, and return-path ellipsometry with special reflectors. In this section, the basic principles and the main features of these approaches will be introduced.

## 3.1 Combination of topometry and ellipsometry

In order to simultaneously determine the topometry and optical constants of surfaces, the combinations of ellipsometry and topometric measurements are proposed, e.g., laser interferometry [16], microscopic fringe projection [15] and white light interferometry [14]. A high numerical aperture (NA) microscope objective is used to collect the reflected light, which is shown in Fig. 3.1. The topometric measurements can measure heights in relative to a plane of reference and ellipsometric measurements can measure the optical constants or film thicknesses. The common feature of these configurations is the off-axis focusing method which can provide tilted irradiation on the surface, high lateral resolution, and collect the reflected light from the nonplanar surface. High NA microscope objectives can measure steep inclinations of surfaces. However, the working distance is short, e.g, an objective with a NA of 0.8 has a working distance of 1 mm.

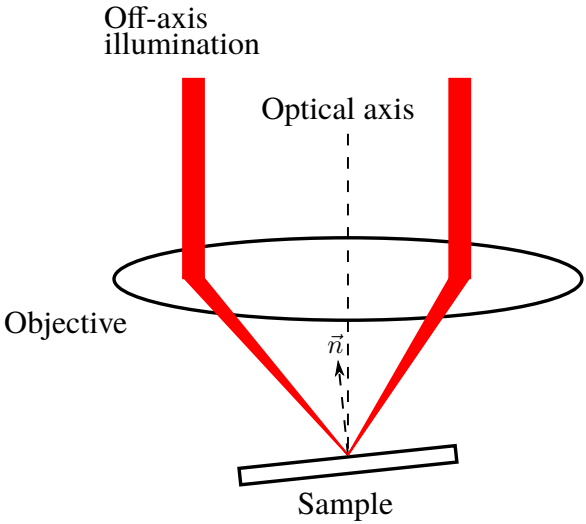



**Figure 3.1**: Internal focusing and off-axis illumination with a tilted sample.

In contrast to the internal focusing, Wirth [21] proposed micro-deflection-ellipsometry to combine topometric and ellipsometric measurements , which is shown in Fig. 3.2. He used a lens system to collect the reflected light for the polarization state generator (PSG), and a beamsplitter to split the reflected light to a position sensitive detector (PSD) and an ellipsometric detector. The PSD can determine the surface orientation and the ellipsometric parameters can be obtained by the ellipsometric detector. In order to receive evaluable signals from the curved surface, the diameter of the first lens should have a large aperture. Therefore, Fresnel lenses are used in the optical system. Compared to the internal focusing, this configuration has a higher working distance of 100 mm.

## 3.2 Polarization model for azimuth deviations

Lee and Chao [13] found the azimuth deviation of the polarizer is the same as the deviation of the surface normal in a calibrated rotating-analyzer ellipsometer. The relationship can be described by Mueller matrices [1]:

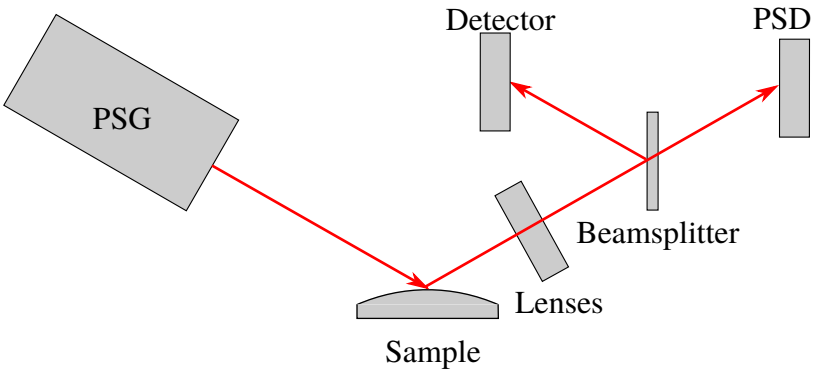



**Figure 3.2**: Combination of topometry (PSD) and ellipsometry (adapted from Wirth [21]).

$$\mathbf{M}_{meas} = \mathbf{M}_A \cdot \mathbf{R}(A) \cdot \mathbf{M}_{sample} \cdot \mathbf{R}(-P) \cdot \mathbf{M}_P, \tag{3.1}$$

where $\mathbf{M}_{meas}$, $\mathbf{M}_{sample}$, $\mathbf{M}_A$, $\mathbf{M}_P$ and $\mathbf{R}$ are the Mueller matrices of the measured matrix, the nondepolarizing isotropic sample, the analyzer, the polarizer and the rotation matrix, respectively. $P$ and $A$ of the rotation matrix $\mathbf{R}$ represent the rotation angle of the polarizer and the analyzer. They used Eq. 3.1 and a three-intensity imaging ellipsometer to measure the surface topography and the coating thickness of a lens.

Neuschaefer-Rube and Holzapfel [19] proposed a method to measure surface geometry and material distribution. They used the internal focusing to collect the reflected beam from the curved surface, which is introduced in section 3.1. Despite of the similar configuration, the surface inclinations can be determined directly by the polarization model without other topometric measurement methods. The polarization model is expressed as:

$$\mathbf{M}_{meas} = \mathbf{R}(\theta_{out}) \cdot \mathbf{R}(-\phi) \cdot \mathbf{M}_{sample} \cdot \mathbf{R}(\phi) \cdot \mathbf{R}(\theta_{in}), \tag{3.2}$$

where $\theta_{in}$, $\theta_{out}$ and $\phi$ are the azimuthal rotation angles on the principle plane of focusing optics. The angle of incidence and the surface orientation can be obtained by the eight-zone-measurement algorithm. After the scanning for the whole surface, the profile can be reconstructed from the surface inclination (gradient data).

Johs and He [12] used a return-path ellipsometer to measure samples which have a wobble effect. The configuration of return-path ellipsometry will be introduced in Fig. 3.3. They established a Mueller matrix model to describe the measurement system. The model is shown as:

$$\mathbf{M}_{meas} = \mathbf{R}(rec) \cdot \mathbf{M}_{sample} \cdot \mathbf{M}_{mirror} \cdot \mathbf{M}_{sample} \cdot \mathbf{R}(src), \tag{3.3}$$

where $rec$ and $src$ are the rotation angles of the receiver and the source. They compensated a $\pm 0.8^\circ$ substrate wobble and reduced the signal variation to less than 2%.

Li et al. [17] considered the effect of the incident plane deviation and proposed a Mueller matrix model to describe the Mueller matrix of the tilt surface as:

$$\mathbf{M}_{meas} = \mathbf{R}(-\alpha) \cdot \mathbf{M}_{sample} \cdot \mathbf{R}(\alpha). \tag{3.4}$$

They believed the two rotation have the same absolute values but different signs. By using Mueller matrix ellipsometry, they successfully measured the oxide layer thickness and the curvature radius for a spherical lens.

Duwe et al. [6] modified the Muller matrix model of Li et al. because of a significant mismatch at larger tilt angles. The modified model is described as:

$$\mathbf{M}_{meas} = \mathbf{R}(-\delta) \cdot \mathbf{M}_{sample} \cdot \mathbf{R}(\gamma), \tag{3.5}$$

where $\delta$ and $\gamma$ are rotation angles of the Mueller rotation matrix. In contrast to the model of Li et al., they assumed the two rotation angles have different signs and values. They used a spectroscopic imaging ellipsometer to measure single-layer coating on a microlens.

## 3.3 Return-path ellipsometry

In return-path ellipsometry, the light beam reflected from the surface is reflected back to the same position from the surface by a mirror [20, 2]. Fig. 3.3 illustrates the schematic of return-path ellipsometry. The advantages of this configuration are simple construction, suitable for process monitoring, and higher sensitivity to the optical properties of surfaces than conventional ellipsometers. Please refer to [3] for more details.

In most semiconductor process, samples usually need to rotate to obtain uniform layers, e.g., plasma-enhanced chemical vapor deposition and epitaxial growth process. The rotation of samples inevitably produces a wobble effect because the rotation axis and the surface normal of the sample are not parallel. As mentioned in section 2, ellipsometry is very sensitive to the angle of incidence and the sample position. In order to obtain accurate measurements, Haberland et al. [9] used return-path ellipsometry and replaced the plane mirror by using a spherical mirror. In geometry ray tracing, every ray which passes the vertex of the spherical mirror is reflected back along the original path. This configuration can effectively reduce the error from the angle deviation for sample rotation and sample wobbling during the manufacturing process.

In order to solve the alignment problem between the sample and the detector, Hartrumpf and Negara [10] developed a laser scanner to overcome this limitation

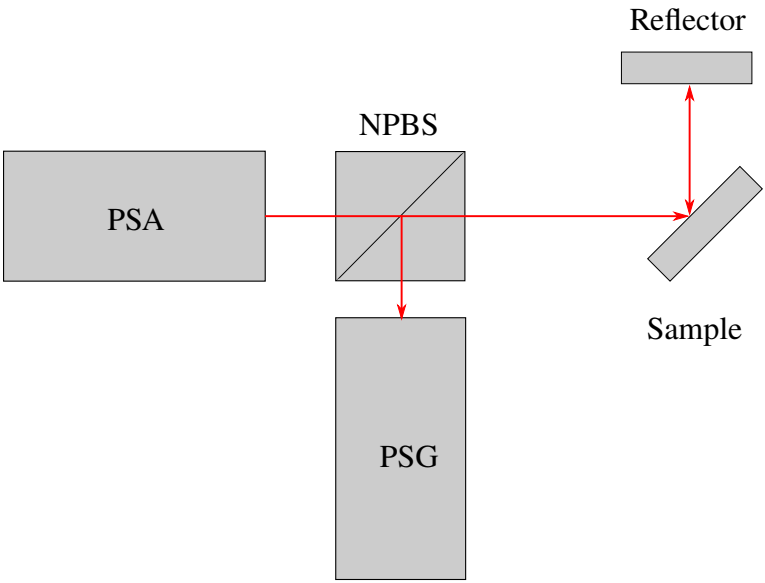



**Figure 3.3**: Configuration of return-path ellipsometry, where PSA, PSG and NPBS are polarization state generator, polarization state analyzer and non-polarizing beamsplitter, respectively.

by a retroreflector (retroreflective sheet). The principle is based on return-path ellipsometry which is shown in 3.3. They used a retroreflector as a reflector. A retroreflector can return the light beam from the sample back on the same beam path with only a phase difference of $180°$. In other words, the polarization effect is the same as an ideal mirror. In this configuration, the alignment condition for the sample and the detector is fulfilled at an angle deviation up to $30°$. Chen et al. used this concept to develop a ellipsometer and measured the ellipsometric parameters and the refractive index for nonplanar surfaces [4, 5].

# 4 Discussion and comparison

For ellipsometric measurements of nonplanar surfaces, combination of topometry and ellipsometry is a straightforward method. This method can achieve very high lateral resolution (about $2.8\ \mu$m in [19]). However, the hardware of topometry will increase the complexity of the whole system, especially for the system alignment and the calibration. In addition, these methods use a focused beam. In order to acquire accurate results, correct focusing planes of the measurement beam are important. Auto focusing methods are applied in these approaches. For the measurement of the whole surface, vertical and xy scanning for every point are necessary, which is very time-consuming.

Polarization models for azimuth deviations provide another solution for surface geometry. This method can be easily applied to conventional ellipsometers without extra hardware. Nonetheless, the range of topometric measurements is limited to a small range because the polarization characteristic of the analyzer (waveplate) is sensitive to the incident angle [8]. Waveplates are constructed by birefringent materials and designed for a normal incident angle. Thus, the retardance of a waveplate will change when the incident angle is not normal. Large incident angles for the waveplate will induce significant errors of the retardance. On the other hand, if the sample is tilted, according to the calculation in 2, the beam offset from the detector is large. Adjustment of the position for the detector is necessary and also time-consuming.

Return-path ellipsometry has a high sensitivity of optical properties of materials due to the double reflection from the sample. Special reflectors (spherical mirror and retroreflector) can achieve ellipsometric measurements for nonplanar surfaces. However, the disadvantages of this configuration are the need for a high power light source and the polarization distortion induced by the non-polarizing beamsplitter. The non-polarized beamsplitter loses a large amount of power of the light source (more than 75%). Moreover, the non-polarized beamsplitter is not an ideal component in polarization optics [18, 22]. Therefore, the calibration of the beamsplitter is necessary.

# 5 Summary

In this report, we have introduced different approaches for ellipsometric measurements of nonplanar surfaces including the principles and the main features. Each approach has its own advantages, disadvantages and suitable application fields. Conventional ellipsometers can only measure samples with flat or nearly flat surfaces. However, there is an urgent need for ellipsometric measurements of nonplanar surfaces in the market, e.g., lens coatings and varnish layer on metallic objects. Further research should be conducted in theory and hardware development for needs of industries.

# References

[1] R. M. A. Azzam and N. M. Bashara. *Ellipsometry and polarized light*. Amsterdam and Oxford: North-Holland, 1977. ISBN: 9780444870162.

[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] Chia-Wei Chen. "An Overview of Return-Path Ellipsometry". 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. 1–10. ISBN: 978-3-7315-0936-3.

[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, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 38.1 (2020), p. 014005. ISSN: 2166-2746. DOI: 10.1116/1.5121854.

[6] 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.

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

[8] Honggang Gu et al. “Study of the retardance of a birefringent waveplate at tilt incidence by Mueller matrix ellipsometer”. In: *Journal of Optics* 20.1 (2017), p. 015401. ISSN: 2040-8986. DOI: 10.1088/2040-8986/aa9b05. URL: https://iopscience.iop.org/article/10.1088/2040-8986/aa9b05/pdf.

[9] K. Haberland et al. “Ellipsometric and reflectance-anisotropy measurements on rotating samples”. In: *Thin Solid Films* 313-314 (1998), pp. 620–624. ISSN: 00406090. DOI: 10.1016/S0040-6090(97)00897-3.

[10] Matthias Hartrumpf and Christian Negara. *Configurable retro-reflective sensor system for the improved characterization of the properties of a sample*. WO/2017/207681. 7.12.2017.

[11] Eugene Hecht. *Optics*. 4th ed. Reading Mass.: Addison-Wesley, 2002. ISBN: 9780805385663.

[12] 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.

[13] 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.

[14] K. Leonhardt. “Ellipso-Height Topometry, EHT: Extended topometry of surfaces with locally changing materials”. In: *Optik - International Journal for Light and Electron Optics* 112.9 (2001), pp. 413–420. ISSN: 0030-4026. DOI: 10.1078/0030-4026-00079.

[15] K. Leonhardt, U. Droste, and H. J. Tiziani. “Topometry for locally changing materials”. In: *Optics Letters* 23.22 (1998), p. 1772. ISSN: 1539-4794. DOI: 10.1364/OL.23.001772.

[16] K. Leonhardt, H.-J Jordan, and H. J. Tiziani. “Micro-Ellipso-Height-Profilometry”. In: *Optics Communications* 80.3-4 (1991), pp. 205–209. ISSN: 00304018. DOI: 10.1016/0030-4018(91)90251-8.

[17] 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.

[18] Yeng-Cheng Liu, Yu-Lung Lo, and Chia-Chi Liao. “Compensation of non-ideal beam splitter polarization distortion effect in Michelson interferometer”. In: *Optics Communications* 361 (2016), pp. 153–161. ISSN: 00304018. DOI: 10.1016/j.optcom.2015.09.099.

[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] Frank Wirth. “Zur Erfassung von form- und materialbedingten Oberflächenstrukturen mit Mikro-Deflexions-Ellipsometrie”. Kassel., Univ., Diss., 2008. PhD thesis. 2008. URL: https://kobra.bibliothek.uni-kassel.de/handle/urn:nbn:de:hebis:34-2008020120181.

[22] Song Zhang et al. “Characterization of beam splitters in the calibration of a six-channel Stokes polarimeter”. In: *Journal of Optics* 20.12 (2018), p. 125606.

# Multimodal 3D Semantic Segmentation

*Fabian Duerr*

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

Technical Report IES-2019-06

## Abstract

Understanding and interpreting a scene is a key task of environment perception for autonomous driving, which is why autonomous vehicles are equipped with a wide range of different sensors. Semantic Segmentation of sensor data provides valuable information for this task and is often seen as key enabler. In this report, we're presenting a deep learning approach for 3D semantic segmentation of lidar point clouds. The proposed architecture uses the lidar's native range view and additionally exploits camera features to increase accuracy and robustness. Lidar and camera feature maps of different scales are fused iteratively inside the network architecture. We evaluate our deep fusion approach on a large benchmark dataset and demonstrate its benefits compared to other state-of-the-art approaches, which rely only on lidar.

## 1 Introduction

One of the key challenges of autonomous driving is the understanding of the vehicle's environment. Therefore, autonomous vehicles are equipped with a wide range of sensor modalities, usually including, camera, lidar, radar and ultrasonic

sensors. With different complementary sensors available, shortcomings of an individual sensor type can be compensated by other sensor types, increasing accuracy and robustness. In this work, we focus on camera and lidar sensors.
Understanding and interpreting a scene is a key task of environment perception for autonomous driving, which makes semantic segmentation of sensor data valuable. For camera images, assigning a class label to every image pixel has been addressed very successfully with Convolutional Neural Networks (CNNs) over the past years, achieving impressive results on road and urban scenes [5]. When dealing with 3D lidar point clouds however, the first challenge is a proper representation, enabling the application of CNNs. One possibility is the lidar's native range view, which has shown promising results [15, 16]. This allows the application of established image segmentation architectures.
Having different sensors available with an overlapping field of view, allows for approaches that fuse the data of different sensors to improve the robustness and overall accuracy. When addressing the fusion of camera and lidar data, some challenges arise. One is a substantial difference in their resolution and another is their considerable difference in measurements and sensor space. While a camera observes brightness values resulting in an image, a lidar measures the distance to its environment, generating a sparse 3D point cloud. Additionally, different fusion strategies must be considered. Following [4], these are the fusion of the sensor data (early fusion), the fusion of the predictions for lidar and camera data (late fusion) or the fusion of the featues maps inside a CNN (deep fusion). In this work, we propose a deep fusion approach, applied to the range view representation, which makes use of camera and lidar data to calculate a semantic segmentation of lidar point clouds. The contributions of this work are twofold:

- First, we propose a fusion module, which takes camera and lidar features, transforms them into a common space and fuses them afterwards.
- Second, we propose a fusion architecture building upon the fusion modules and apply them iteratively throughout our network, following the idea of iterative deep aggregation [26]. This way, we are able to fuse aggregated features of both sensors at different scales and maximize the fused information

# 2 Related work

## 2.1 2D Semantic segmentation

The success of deep learning applied for scene parsing and semantic segmentation [13, 21, 8] is closely related to its success in classical image classification [22, 10, 7]. One widely used approach are Fully Convolutional Neural Networks (FCNN) [13], which calculate a pixel-wise prediction for a given image in an end-to-end fashion. [13] replaced the fully connected layers of common classification architectures with 1x1-convolutions, thereby replacing the original image classification with a pixel-wise classification.
One main challenge, recent works have focused on, is the loss of spatial resolution while aggregating information. It is of great importance to capture the global context of a scene as well as fine local structures. DeepLabv3 [3, 2] addresses this by 'atrous' convolutions, which increase the size of the receptive fields without reducing resolution or increasing filter sizes. 'Atrous' convolutions with different rates are employed in parallel to exploit context at different scales. In [26], an aggregation architecture is presented, which the authors call deep layer aggregation (DLA), also targeting the challenge of extracting meaningful semantic features while preserving spatial information. PSPNet [29] combines local and global context by a pyramid pooling module, which aggregates the global context at different scales and appends it to the original feature maps. OcNet [27] adapts the idea of the pyramid pooling module and multiscale 'atrous' convolutions by introducing an object context module, which exploits object context at different scales, instead of spatial context.

## 2.2 3D Semantic segmentation

When addressing semantic segmentation of 3D point clouds with CNNs, the first thing to consider is the representation of the point clouds. In recent works, multiple different representations are proposed. PointNet [18] uses the raw and unstructured point clouds directly as input by applying pointwise 1x1-convolutions and a symmetric operation for feature aggregation. Because a single global feature aggregation limits the ability to capture spatial relations, the

authors proposed PointNet++ [19], which applies individual PointNets to local regions and aggregates the resulting local features in a hierarchical fashion. [23] converts the point clouds into a voxel grid and applies a 3D-FCNN, followed by a Conditional Random Field (CRF) to refine the results. A bird's eye view (BEV) with the vertical axis as feature channel is used by [28] to retrieve a 2D representation of the point clouds. Having a 2D representation, they're using the U-Net architecture [21], known from image segmentation.
When working with point clouds generated by a lidar sensor, the range view is another possibility of representation. SqueezeSeg [24] was one of the first works using the range view for a segmentation task. Their goal was the segmentation of road objects, with an improved version released in [25]. Another approach is RangeNet++ [16], which employs the DarkNet53 backbone [20] for full semantic segmentation. [14] proposed LaserNet, which uses the range view as input for object detection, while one of their intermediate results is a semantic segmentation of the input. Their architecture is based on deep layer aggregation. Transforming the point cloud into its range view and applying established 2D image segmentation architectures mostly outperforms other forms of representations while being faster. Therefore, our work also builds upon the range view representation.

## 2.3 Multimodal 3D semantic segmentation

Multi-sensor fusion architectures using camera and lidar mostly focus on object detection [4, 17, 11, 12, 15]. Only [15] also tackles the task of 3D semantic segmentation, using the range view as input representation. Camera image feature maps, extracted by three ResNet blocks [7], and extracted lidar feature maps from the range view are concatenated and passed to a LaserNet, which serves as DLA for the semantic segmentation. In contrast to applying early fusion and fusing the RGB values with the range view, this approach aggregates camera image information first, using the original usually much higher resolution of the camera image. This deep fusion allows for more information being preserved and exploited for the semantic segmentation of the lidar point cloud. While considerably improving the mean Intersection over Union over all classes (mIoU) on distant content $(+5.19)$, the overall improvements are rather small $(+0.25)$.

We're also using deep layer aggregation and the full camera image resolution for deep fusion of camera and lidar. In contrast to [15], which fuses the features before applying their DLA network (LaserNet), we're applying a DLA network to both, the lidar range view and the camera image, separately but fuse both networks following iterative deep aggregation [26]. As a result, our deep fusion approach is able to aggregate and use more information from the camera for the semantic segmentation of the lidar point cloud.

# 3 Iterative deep fusion and aggregation

In this section, we present our range view input representation, our fusion module and the network architecture, used for the fusion of the lidar and camera input.

## 3.1 Range view

Commonly used lidar sensors usually observe their environment by spinning a set of vertically stacked lasers around their vertical axis. The position of a laser in this stack is often referred to as channel, corresponding to an elevation angle. The Velodyne HDL-64E, used to record the SemanticKitti dataset [1, 6], has 64 channels, an azimuth resolution of approximately $0.17°$ and an elevation resolution of $^1/_3{}^\circ$ for the upper and $^1/_2{}^\circ$ for the lower half of the lasers. The sensor provides measurements $\boldsymbol{o}_i = (c_i, \phi_i, r_i, e_i)$, with channel id $c_i$, azimuth angle $\phi_i$, measured distance $r_i$ and reflectance $e_i$. The corresponding 3D points are

$$\boldsymbol{p}_i = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} r_i \cos(\theta_i) \cos(\phi_i) \\ r_i \cos(\theta_i) \sin(\phi_i) \\ r_i \sin(\theta_i) \end{pmatrix}, \tag{3.1}$$

omitting correction factors. The elevation angle $\theta_i$ is derived from the sensor configuration and the channel id $c_i$.
We generate a range view by mapping every point or measurement to a row and column index. Having measurements from a Velodyne HDL-64E, the row and column indices are calculated by using the channel as row index and discretizing

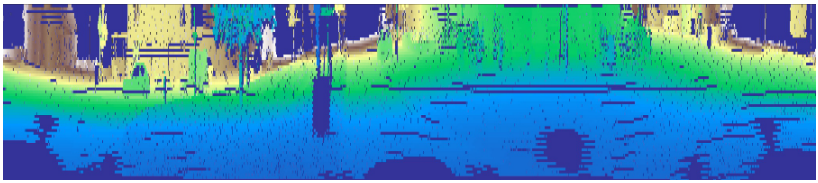


**Figure 3.1**: range view showing the lidar depth measurements.

the azimuth angle. If only the 3D points $\boldsymbol{p}_i$ are provided, the azimuth and elevation angle are given by

$$\phi_i = -\text{arctan2}(y_i, x_i) \quad \text{and} \quad \theta_i = \arcsin\left(\frac{z_i}{r_i}\right). \tag{3.2}$$

Finally, for a range view resolution of $h \times w$, the image coordinates $\boldsymbol{u}_i^{\text{li}} = (u_i^{\text{li}}, v_i^{\text{li}})$ are

$$u_i^{\text{li}} = \begin{cases} \left\lfloor 0.5 \cdot h \cdot \frac{\theta_i - \theta_{\text{up}}}{\theta_{\text{mid}} - \theta_{\text{up}}} \right\rfloor & \theta_i \geq \theta_{\text{mid}} \\ \left\lfloor 0.5 \cdot h \cdot \left(1 + \frac{\theta_i - \theta_{\text{mid}}}{\theta_{\text{down}} - \theta_{\text{mid}}}\right) \right\rfloor & \theta_i < \theta_{\text{mid}} \end{cases}, \tag{3.3}$$

$$v_i^{\text{li}} = \left\lfloor 0.5 \cdot \left(1 + \frac{\phi_i}{\pi}\right) \cdot w \right\rfloor, \tag{3.4}$$

with a vertical field of view $\theta_{\text{fov}} = \theta_{\text{up}} - \theta_{\text{down}} = 2° - (-24.8°) = 26.8°$ and the border angle between the two vertical resolutions $\theta_{\text{mid}} = -{}^{26}/_{3}{}^{\circ}$. Following this, we're mapping the input measurements $r$, $e$, $x$, $y$ and $z$ to the 2D range view, receiving a $5 \times h \times w$ input tensor $\mathbf{R}$. The depth channel $(r)$ is visualized in Fig. 3.1.

Ego motion, uncertainty and non-uniformity of the angles can lead to mapping collisions. As a result, more than one point is mapped to the same range view pixel. This implies not only a loss of information but also missing predictions for the shadowed points. The latter isn't an issue for object detection, for semantic segmentation however, it has to be considered. Therefore, a post-processing step based on the labeled points is required to compute class labels for the shadowed points. Following the simplest one, we assign the same label to all

measurements projected on the same range view pixel. Another approach is based on k-nearest neighbor [16]. We will investigate the post-processing step in future work. In this work, we're focusing on the feature fusion.

## 3.2 Feature transformation and fusion

A crucial part of our work is the feature fusion, which fuses the lidar and camera features. We're choosing the range view as our reference system and project camera features into it. The inverse projection, from lidar to camera, is mathematically given by the equation

$$\begin{pmatrix} u_i^{\text{cam}} \\ v_i^{\text{cam}} \\ 1 \end{pmatrix} = \mathbf{K} \cdot \mathbf{T}_{\text{li2cam}} \cdot \begin{pmatrix} \boldsymbol{p}_i \\ 1 \end{pmatrix}, \tag{3.5}$$

with the camera matrix $\mathbf{K}$ and transformation matrix from lidar to camera $\mathbf{T}_{\text{li2cam}}$. The calculated pixel indices define the correspondence between 3D points and camera pixels. For this correspondence being still valid after scaling the range view by $\beta$ or the camera image by $\alpha$, the following extensions are made

$${}^{\alpha}\boldsymbol{u}_i^{\text{cam}} = \begin{pmatrix} \lfloor u_i^{\text{cam}} \cdot \alpha \rfloor \\ \lfloor v_i^{\text{cam}} \cdot \alpha \rfloor \end{pmatrix} \quad \& \quad {}^{\beta}\boldsymbol{u}_i^{\text{li}} = \begin{pmatrix} \lfloor u_i^{\text{li}} \cdot \beta \rfloor \\ \lfloor v_i^{\text{li}} \cdot \beta \rfloor \end{pmatrix}, \quad \text{with } \alpha, \beta \in [0, 1]. \tag{3.6}$$

Given scalable projection indices, we're now able to project camera features $\mathbf{I}^{\alpha}$ into the range view $\mathbf{R}^{\beta}$, following

$$\mathbf{R}^{\beta}[\,{}^{\beta}\boldsymbol{u}_i^{\text{li}}] \coloneqq \mathbf{I}^{\alpha}[\,{}^{\alpha}\boldsymbol{u}_i^{\text{cam}}]. \tag{3.7}$$

This is a fixed, geometrically motivated mapping, considering only one location per 3D point in the camera feature maps. To capture more context and to compensate errors in the calibration, we apply a learnable function $F_{\boldsymbol{w}}$ before performing the fixed projection, resulting in

$$\mathbf{I}_F^{\alpha} = F_{\boldsymbol{w}}(\mathbf{I}^{\alpha}) \quad \text{and} \quad \mathbf{R}_{\boldsymbol{w}}^{\beta}[\,{}^{\beta}\boldsymbol{u}_i^{\text{li}}] \coloneqq \mathbf{I}_F^{\alpha}[\,{}^{\alpha}\boldsymbol{u}_i^{\text{cam}}]. \tag{3.8}$$

The fusion module shown in Fig. 3.2 builds upon this to implement the camera feature transformation. We're using a 3x3 convolution followed by Batch Norm [9] and ReLu as learnable function $F_{\boldsymbol{w}}$. The projected camera features and the lidar features are concatenated and fused by ResNet blocks.

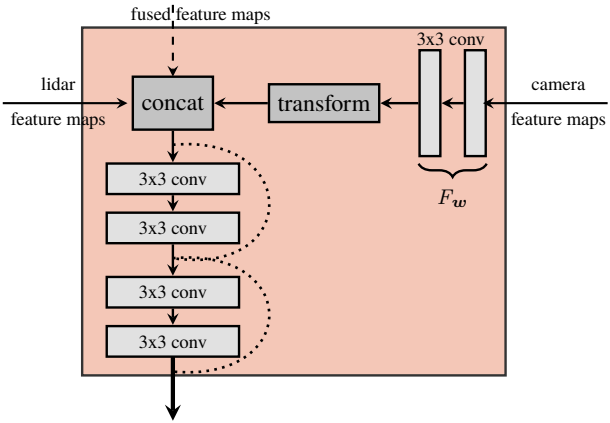



**Figure 3.2**: The main building block of our architecture. The fusion module transforms the camera features into the lidar range view. Afterwards, lidar feature maps, camera feature maps and optionally fused features maps from the stage before are fused.

## 3.3 Network architecture

Our proposed network architecture is shown in Fig. 2.3 and has three main components. First, a DLA network called Lidar-Net (I) for processing the lidar range view and calculating lidar features. It follows the proposed architecture of [14], which itself is based on [26]. By using a DLA architecture, we ensure to efficiently aggregate multi-scale lidar features. The second component is another DLA network (II) with the same architecture for processing the camera image. Additionally, we downsample the camera image before applying the DLA network. The resolution of the camera image is much higher than of the lidar image, so the induced loss in spatial information is small, whereas the aggregated semantic information are considerably improved. We follow the ResNet architecture and downsample the camera image with a strided convolution and max pooling by a factor of four. This also decreases the run time and memory requirements. The last component are fusion blocks (III), which apply the previously presented feature transformation and fusion. They

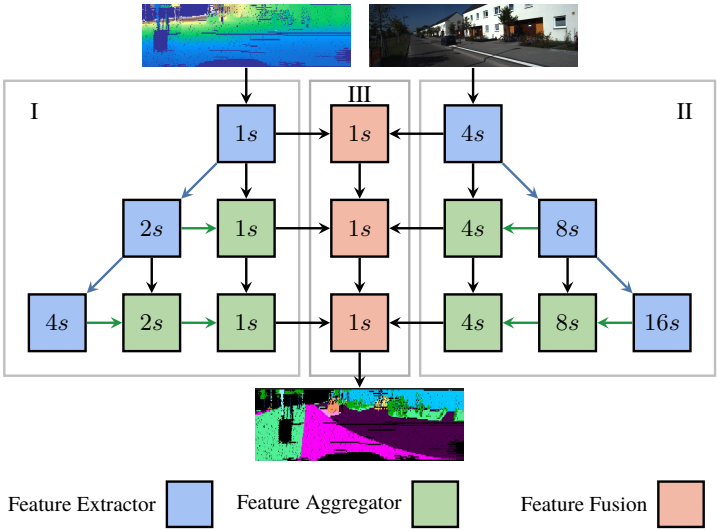



**Figure 3.3**: Our proposed fusion architecture, which fuses the lidar and camera features iteratively, following the idea of iterative deep aggregation [26]. The labels indicate the output stride of the individual blocks. We use the same network parameters for (I) and (II) as [14].

follow the idea of a feature aggregator except that they transform and aggregate features of different sensors instead of different scales of one sensor.

# 4 Experiments

## 4.1 SemanticKitti

We're evaluating our approach on the SemanticKitti dataset [1, 6], which contains labels for 19 classes for the single scan benchmark. A total of 22 labeled sequences results in 43552 labeled scans. The official split allocates sequences 0-10 for training and sequences 11-21 for testing, for which the labels haven't been published. However, the official benchmark doesn't support the usage of the camera images, meaning for our evaluation, only the sequences

with published labels 0-10 can be used. Therefore, were excluding sequences 02, 06 and 10 from training and validation and use them only in the end for testing. This results in 6963 frames for testing and 16238 for training and validation. We follow the official evaluation metric and report the mean Intersection-over-Union (mIoU). For our approach, only the lidar scan parts overlapping with the camera's field of view in the front of the car can be used.

## 4.2 Implementation details

Our training starts with an initial learning rate of $10^{-4}$, which is then multiplied in each training iteration it by $10^{\frac{-2 \cdot \text{it}}{\text{it}_{\text{max}}}}$. Thereby, the learning rate exponentially decreases by $1/100$ during training. We train our network for $50k$ iteration with a batch size of 40. To improve generalizability and reduce overfitting, we're using random crops of the whole $360°$ lidar scan for training the lidar net. Although the crop is random, it follows the constraint, that the overlapping field of view with the camera has to be fully inside the crop of size $64 \times 1536$. The fusion modules finally crop the resulting lidar feature maps exactly to the overlapping field of view. Additionally, we apply random flipping horizontally to the lidar and camera images.
To counteract the class imbalance, we're using a class-balanced cross entropy loss for the final output as well as the auxiliary loss. The latter is used on the final feature map of the Lidar-Net. Following the proposed settings of PSPNet [29], we're weighting the auxiliary loss by $0.4$

## 4.3 Results

We evaluate our approach and present the improvements gained by the fusion of lidar and camera image features. Therefore, we compare the results of our deep fusion architecture, called Fusion-Net, to Lidar-Net, which uses only the lidar scans. The results of both approaches are shown in Tab. 4.1. Overall, our fusion approach outperforms Lidar-Net by a considerable margin, and also the majority of the individual classes considerably benefit from the deep fusion approach.

| Approach | road | sidewalk | parking | other-ground | building | car | truck | bicycle | motorcycle | other-vehicle |
|---|---|---|---|---|---|---|---|---|---|---|
| Lidar-Net | 93.1 | 76.8 | **56.1** | **3.4** | 67.1 | 81.7 | **42.0** | 23.2 | 39.8 | 29.0 |
| Fusion-Net | **93.2** | **77.0** | 55.9 | 0.4 | **74.0** | **82.0** | 37.8 | **26.4** | **43.1** | **29.1** |

| Approach | vegetation | trunk | terrain | person | bicyclist | motorcyclist | fence | pole | traffic sign | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|
| Lidar-Net | 78.0 | 58.1 | 67.2 | 35.6 | **11.8** | **2.4** | 57.2 | 36.4 | 39.9 | 47.3 |
| Fusion-Net | **81.4** | **65.8** | **72.0** | **42.7** | 11.0 | 0.3 | **59.4** | **49.6** | **45.6** | **49.8** |

**Table 4.1**: Comparison of the results of our deep fusion architecture and the purely lidar based Lidar-Net

# 5 Conclusion and Outlook

In this work, we've presented a deep learning approach for semantic segmentation of 3D lidar point clouds. Our approach uses a range view representation of the lidar scans, enabling the application of established image segmentation approaches. Furthermore, we use camera image feature maps of different scales and iteratively fuse them inside our network with the lidar feature maps. Our experiments underline the advantages of our deep fusion approach, which outperforms a lidar-only approach by a considerable margin in terms of the mIoU. Also, most of the individual classes considerably benefit from the fusion. For the future, we plan to further improve our fusion modules and thereby increase the benefits of our fusion architecture. We're also planning a more in depth analysis of the benefits of fusing camera and lidar data for 3D semantic segmentation.

# References

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

[2] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2016), pp. 834–848.

[3] Liang-Chieh Chen et al. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”. In: *International Conference on Learning Representations (ICLR)* (2015).

[4] Xiaozhi Chen et al. “Multi-view 3D Object Detection Network for Autonomous Driving”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 6526–6534.

[5] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3213–3223.

[6] 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), pp. 3354–3361.

[7] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 770–778.

[8] Kaiming He et al. “Mask R-CNN”. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[9] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *ArXiv* abs/1502.03167 (2015).

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[11] Jason Ku et al. “Joint 3D Proposal Generation and Object Detection from View Aggregation”. In: *IEEE International Conference on Intelligent Robots and Systems (IROS)* (2017), pp. 1–8.

[12] Bo Li, Tianlei Zhang, and Tian Xia. “Vehicle Detection from 3D Lidar Using Fully Convolutional Network”. In: *ArXiv* (2016).

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3440.

[14] Gregory P. Meyer et al. “LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving”. In: *ArXiv* abs/1903.08701 (2019).

[15] Gregory P. Meyer et al. “Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

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

[17] Charles Ruizhongtai Qi et al. “Frustum PointNets for 3D Object Detection from RGB-D Data”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 918–927.

[18] 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), pp. 77–85.

[19] Charles Ruizhongtai Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *NIPS*. 2017.

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

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* Vol.9351 (2015), pp. 234–241.

[22] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[23] Lyne P. Tchapmi et al. “SEGCloud: Semantic Segmentation of 3D Point Clouds”. In: *International Conference on 3D Vision (3DV)* (2017), pp. 537–547.

[24] 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), pp. 1887–1893.

[25] 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), pp. 4376–4382.

[26] Fisher Yu et al. “Deep Layer Aggregation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2403–2412.

[27] Yuhui Yuan and Jingdong Wang. “OCNet: Object Context Network for Scene Parsing”. In: *ArXiv* abs/1809.00916 (2018).

[28] 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), pp. 399–408.

[29] Hengshuang Zhao et al. “Pyramid Scene Parsing Network”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 6230–6239.

# Part Affinity Field based Activity Recognition

*Thomas Golda*

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

Technical Report IES-2019-11

## Abstract

This report presents work and results on Activity Recognition using Part Affinity Fields for real-time surveillance applications. Starting with a short introduction to the motivation, this report gives a detailed overview over the key idea of the pursued approach and explains the basic ideas. In addition a variety of experiments on various subjects are presented, like i) the impact of the number of input frames, ii) the impact of different simple dimensionality reduction approaches, and iii) a comparison on how multi-class and binary problem formulation influence the performance.

## 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 is 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. Recent progress in the fields of classification, pattern recognition and time series prediction has also brought the field of activity recognition into the focus of application-oriented research for surveillance scenarios. This report proposes a new human pose based approach on classifying the behavior of pedestrians. It presents details of the architecture and various experiments on different considered time horizons as well as various considerations that were conducted in order to tackle the problem of activity recognition in such scenarios. The presented approach is preparatory work for future activities on human-centered abnormal behavior detection.

# 2 Part affinity fields for activity recognition

## 2.1 Human pose estimation in the wild

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. Usually the used sensors are classical video cameras or depth cameras delivering RGB or RGB-D information respectively. This work focuses on the 2D case using classical cameras and RGB data. This decision is driven by the corresponding problem domain, namely video surveillance in urban setups, where typical camera setups consist of RGB cameras. To this point, RGB-D cameras are rarely used. As a consequence, the resulting skeletons produced by human pose estimation algorithms consist of keypoints in a two-dimensional space with $V \subset \mathbb{R}^2$.

## 2.2 Part affinity fields

For this approach, the framework *OpenPose* by Cao et al. [1], which belongs to the group of *bottom-up* methods, is used. Contrary to *top-down* methods, bottom-


**Figure 2.1**: Based on the Part Affinity Fields provided by the OpenPose framework body part maps encoding the presence of a body part are constructed. This is done per frame. The resulting body parts $\mathcal{B}$ for the highlighted person are shown in the lower part of the figure. For visualization purposes the body parts are merged into a single layer image to show that they form an understandable representation of pedestrians.

up methods first locate all keypoints in a given image, which are connected in a subsequent step. To do so, the method proposed in [1] computes *Part Affinity Fields* (PAF) that are used to connect estimated keypoints by adding further semantic information about visible body parts. In detail, the computed PAFs are used for constructing a bipartite keypoint graph that is subject to the final optimization problem which is solved using the Hungarian Method [7].

## 2.3 Architecture

Since the aim of activity recognition in surveillance scenarios is to have a near real-time processing of video footage, typical activity recognition frameworks are not applicable due to their large network architectures and resulting strong hardware requirements. As a result, the focus of this work lies on developing an approach using a much smaller neural network. For the task of image classification

Howard et al. [3] proposed a network architecture called MobileNet that is much smaller, i.e. much less parameters, than most competing architectures achieving comparable performance at the same time. Due to the promising performance of MobileNet for the classification of single images the decision was made to adapt it for an custom and fast activity recognition approach. This choice brings a constraint into the considerations: Working on raw 2D keypoint coordinates is not possible with the pre-defined architecture. However, two possibilities come up when dealing with this problem. The first way is to transform the keypoint coordinates into images, the second to use the body representation already provided by the OpenPose framework. Since it is easy to obtain the latter and is available out of the box when using OpenPose the decision was made to adapt the Part Affinity Fields instead of the raw coordinates.

### 2.3.1 From part affinity fields to human body parts

In order to reduce the number of inputs semantically corresponding Part Affinity Fields $F_{\text{part}} = (F_{\text{part},x}, F_{\text{part},y})$ are aggregated to five body parts: torso, left arm, right arm, left leg and right leg. The following equation shows the formula for computing the corresponding body part using the Part Affinity Fields related to the left leg.

$$\mathcal{B}_{\text{leftLeg}} = \sqrt{\lambda \cdot (\mathcal{F}^2_{\text{leftCalf,x}} + \mathcal{F}^2_{\text{leftCalf,y}})} + \sqrt{\lambda \cdot (\mathcal{F}^2_{\text{leftThigh,x}} + \mathcal{F}^2_{\text{leftThigh,y}})} \quad (2.1)$$

where $\lambda \in \mathbb{R}^+$ is a scaling factor. Note that $\mathcal{B}$ encodes the presence of a body part rather than its direction since this information is lost by transforming the Part Affinity Fields into body parts. However, since the information about single body parts is still available and no further reducing operations are performed, it is still possible to infer the orientation of the represented person.

## 2.4 Training dataset

The decision to use Part Affinity Fields as input to the model architecture makes it impossible to use a pre-trained network since the input volume has five instead

of three channels. Furthermore, the input channels do not correspond to typical structures like they can be encountered in RGB images. Therefore the chosen architecture has to be trained from scratch.
In order to train the network three existing datasets were merged: INRIA Xmas Motion Acquisition Sequences (IXMAS) [11], UT-Interaction [10] and IOSB Multispectral Action Dataset [2]. All these datasets come with different sets of activities $A_{\text{IXMAS}}$, $A_{\text{UT-Interaction}}$ and $A_{\text{IOSB}}$. For this work, the available activities were merged to get a total of 19 activities from all three datasets. About 80% of the available activities come from IXMAS, which is the most diverse dataset of these three. Another reason to chose these datasets is motivated by the viewing perspective and the size of the urban outdoor environment, which showed the best fit to the field of application.
Since all datasets consist of video sequences it is possible to make use of temporal information, which would be typical done by tracking pedestrians. For the initial setup no tracking is considered for performance reasons as tracking of multiple targets would introduce further expensive computations. However, an alternative way to benefit of the available temporal information is inspired by the *anchor cuboids* used in [4]. A schematic overview over this principle is given in Figure 2.2. Given a bounding box $B_t$ at timestep $t$ and a window size $k$ an input volume is constructed by simply aggregating the spatial information at the location of $B_t$ over the last $k$ timesteps

$$\tilde{B}_{t,k} = B_{t-k+1} \oplus ... \oplus B_{t-2} \oplus B_{t-1} \oplus B_t \tag{2.2}$$

where $\oplus$ describes the concatenation operation along the channel axis. Note that each bounding box $B_i$ contains spatially corresponding information from all five body part channels and hence can be written as

$$B_i = (\mathcal{B}_{\text{leftLeg}}, \mathcal{B}_{\text{leftArm}}, \mathcal{B}_{\text{torso}}, \mathcal{B}_{\text{rightArm}}, \mathcal{B}_{\text{rightLeg}}) \tag{2.3}$$

### 2.4.1 Multi-class approach

As mentioned in the introductory part of this section the used dataset consists of 19 activity classes with sequences taken from three different public datasets. The

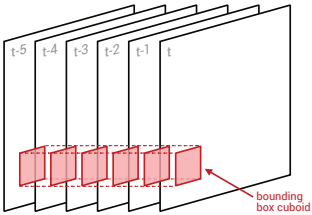



**Figure 2.2**: In order to avoid further computation overhead by performing tracking of pedestrians, temporal cuboids are build as shown in red. For a given timestamp $t$, a given window size $k$ and a bounding box $B_t$ we merge the content of the corresponding bounding boxes $B_{t-k+1}$, ..., $B_t$. All bounding boxes share the same location at different points in time.

set of available activities ranges from everyday activities like *walking*, *sitting down* up to more unusual ones like *kicking* and *punching*. In the multi-class approach every movement of a pedestrian is classified into one of the available classes. For the training of the network the broadly utilized Adam optimizer [5] was used with an initial learning rate of $10^{-3}$. As the training objective, cross-entropy was chosen as it is the most common loss for classification tasks. The learning rate is reduced by a factor of $0.1$ after each 20 epochs without improvement on the used validation set. The whole training was conducted on an Nvidia DGX-1 using a single Tesla V100 card with 32 GiB of memory. This allows to use large batches with around 500 samples per batch. The actual number of samples contained in a single batch is chosen empirically and depends on the regarded number of timesteps $k$. Figure 2.3 illustrates the overall setup of the final architecture, which takes as input a set of $k$ subsequent body part sets of a given person detection $\tilde{B}_{t,k}$. The input is then processed by the adapted MobileNet model and classified as one of the 19 regarded activity classes.

### 2.4.2 Binary approach

In addition to the multi-class approach, a binary classification task with the aim to distinguish between target activities and non-target activities was investigate. As target activities a subset of activities, namely *kick*, *punch*, *hit* and *push* were chosen, since they show quite similar and relevant activities. To encode these

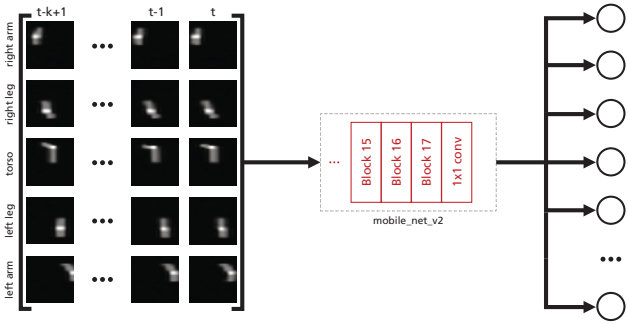



**Figure 2.3**: Input to the model is a volume of size $224 \times 224 \times 5k$ where $k$ is the number of timesteps that are taken into account. Body parts are stacked along the third axis over time. Each detected pedestrian and its corresponding body part maps are resized to $224 \times 224$ as expected by the MobileNet [3] architecture. Every row on the left corresponds to one of the five body parts $\mathcal{B}_i$, each column to a considered timestep. The number of neurons in the final classification layer is furthermore changed to the number of activity classes $c \in \{2, 19\}$.

two classes, the number of output neurons was reduced to two neurons in the model architecture.

A consequence of transforming the multi-class to a binary classification problem is the accentuation of the imbalance between the regarded classes. To address this problem, the imbalance was considered implicitly by changing the used training loss. For this reason the Focal Loss [8] was adapted, which is an extension of the classical cross-entropy loss that introduces a weighting of samples based on the quality of their already achieved classification result.

$$\mathcal{L}_{\text{focal}}(p_t) = -(1 - p_t)^{\gamma} \cdot \log(p_t) \qquad \text{with} \quad \gamma \in \mathbb{R}_0^+ \tag{2.4}$$

As can be seen in the equation above, the difference to the cross-entropy loss comes from an additional factor $(1 - p_t)^{\gamma}$ that reduces the loss for well-classified samples ($p_t > 0.5$). The introduced variable $\gamma \in \mathbb{R}_0^+$ controls how strong the influence of the well-classified samples to the overall loss can get. The higher the value of $\gamma$ the more the samples on which the regarded model already achieves good results affect the computed gradients and hence the training process.

**Figure 3.1**: In order to evaluate the impact of the presented considerations on the overall performance of the approach a dataset was created using two cameras mounted in different heights pointing to the same location that is only used for evaluation.

# 3 Evaluation

## 3.1 IOSB-Ka dataset

For the evaluation an eligible dataset was created that shares many properties with typical public surveillance scenarios. The dataset was recorded at the Fraunhofer Institute for Optronics, System Technologies and Image Exploitation IOSB in Karlsruhe using a common video surveillance setup. It consists of six sequences with an average duration of 56 seconds from two different cameras both showing the same location. The cameras were mounted with different orientations and at different heights. Each video sequence shows a group of people performing actions from a predefined action set that comprises actions like *kicking*, *punching* and *waving*. Figure 3.1 shows two randomly selected frames each taken from one of the two cameras.

## 3.2 Temporal window size

The first part of our experiments were conducted to examine the influence of the temporal window on the overall performance. Since all sequences from our training dataset were recorded with frame rates between 25 and 30 frames per second, the number of consecutive regarded frames has to be chosen long enough to capture the important part of an action. Therefore a series of experiments was performed for values of $k \in \{1, 6, 10, 14, 18, 24\}$. As stated earlier the input

**Table 3.1**: The average precision on the evaluation dataset for the provided activities and the overall mean average precision vary slightly for different values of $k$. The corresponding values are indicated by the identifier PBAR-MF$k$. The model without temporal information, i.e. $k = 1$, is referred to as PBAR-SF. The best results were achieved with a windows size of $k = 10$.

| | **kick** | **punch** | **wave** | **mAP** |
|---|---|---|---|---|
| PBAR-SF | 0.32 | 0.37 | 0.42 | 0.37 |
| PBAR-MF6 | 0.34 | 0.37 | 0.51 | 0.40 |
| PBAR-MF10 | **0.35** | 0.42 | **0.66** | **0.48** |
| PBAR-MF14 | 0.32 | 0.39 | 0.52 | 0.41 |
| PBAR-MF18 | 0.33 | 0.40 | 0.64 | 0.46 |
| PBAR-MF24 | 0.30 | **0.45** | 0.59 | 0.45 |

to the model is a sequence of $k$ consecutive 5-tuples and can be seen as five sequences showing the temporal behavior of different body parts. Table 3.1 shows the results for different window sizes. It is obvious that all approaches perform almost identical for both activities *kick* and *punch*. Even for a time window of almost a second ($k = 24$) the results do not improve. However, the results for *wave* are better and the benefit of including temporal information is clearly visible. The reason for the results on the first two activities might be due to very similar motions in the training dataset and the far wider variety of forms for the same activity in the test set. Another explanation could be, that the model could not learn to distinguish between similar activities. This has to be investigated in future with additional experiments.

## 3.3 Dimensionality reduction

A major drawback that comes up when increasing the number of timesteps and hence the size of the input volume to the neural network is the rising computational complexity at training as well as at inference time. While the first is not too much of a problem, the latter means a direct effect on the frame rate and hence on the ability to be close to real-time. As a consequence the question comes up, whether a reduction of dimensionality achieves comparable performance

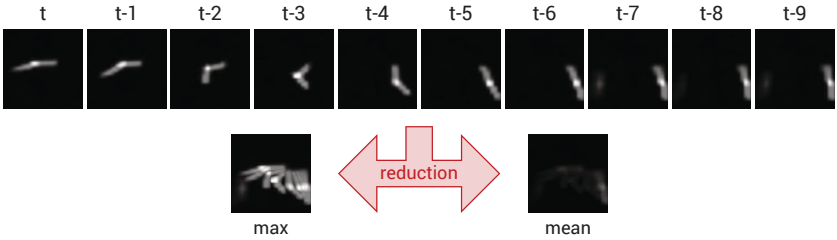



**Figure 3.2**: Given a timestep $t$ and a corresponding input volume $\tilde{B}_{t,k}$, each body part of the input volume is reduced to a single channel. The illustration shows the result of the reduction exemplary for the right leg $\mathcal{B}^i_{\text{rightLeg}}$, where $i$ stands for the corresponding offset to the current timestep. Given $k$ timesteps, the chosen reduction function $f_{\text{reduce}}(\{\mathcal{B}^i_{\text{rightLeg}} \mid i \in \{0, ..., k-1\}\})$ is applied merging on each pixel along the time axis. It is obvious that the *max*-reduction is much more comprehensive for the human, however, the *mean*-reduction does not use any relevant information. It keeps information about the velocity of the action through the amplitude of the output signal.

to the full temporal information. Hence, the network was trained on merged inputs using two ways to reduce the dimensionality: *max* and *mean* reduction. Furthermore, the decision was made to keep the spatial information and perform reduction just over time dimension, i.e. fusing just the information corresponding to the same body part. Figure 3.2 illustrates how the dimensionality reduction works in principle and shows the corresponding outcome for our two considered reduction functions.

**Table 3.2**: In order to investigate the effect of dimensionality reduction two simple approaches were applied to the best performing model PBAR-MF10: *max* and *mean* reduction. In both cases the resulting reduced input volumes do not carry enough information, so that the performance of the trained model drops significantly. The last row also provides results for the non-reduced model as comparison.

| | **kick** | **punch** | **wave** | **mAP** |
|---|---|---|---|---|
| max | 0.29 | 0.29 | 0.57 | 0.38 |
| mean | 0.28 | 0.33 | 0.57 | 0.40 |
| *without* | **0.35** | **0.42** | **0.66** | **0.48** |

Applying reduction to the input decreases training time approximately about 80% from 29.5 minutes to 5.7 minutes per epoch. This effect cannot be observed during inference time where the overall processing time stays identical. A possible reason for this might be that the training is performed using Python. Python is not optimized for memory efficiency and hence moving data from RAM to VRAM might be a bottleneck. The final productive system is written in C++ using the libtorch library provided by the PyTorch development team, which seems to work more efficient when shifting data between devices. However, Table 3.2 indicates that in both cases the performance decreases significantly by a similar amount when using these simple reduction mechanisms.

Since *mean* and *max* reduction appeared to be too strict approaches according to the results presented in Table 3.2, further investigations on the impact of dropping intermediate frames in order to reduce the input size were performed. The dropping is performed in an equally spaced manner using an offset $s \in \mathbb{N}$ and hence results in timesteps $t, t-s, t-2s, ..., t-(k-1)\cdot s$. Written in a more compact way, a sequence of $k$ timesteps with an offset $s$ consists of ordered samples $B_{t-i}$ with $i \in I_{k,s}$ and

$$I_{k,s} = \{\ i \in \mathbb{N}_0 \mid i \leq (k-1)\cdot s\ \wedge\ \exists\, m \in \mathbb{N}_0 : i = m \cdot s\ \} \tag{3.1}$$

This method is referred to as *striding*. Furthermore, for given $k, s \in \mathbb{N}$ the function $\kappa(k, s)$ describes the *sampling window size*.

$$\kappa(k, s) = (k-1)\cdot s + 1 \quad \forall\, k, s \in \mathbb{N} \tag{3.2}$$

For $s = 1$ this resembles the original set of timesteps $t, ..., t-k+1$ for a given *input window size* $k \in \mathbb{N}$. Hence, the sampling window size $\kappa$ equals the input window size $k$. Table 3.3 shows results for a stride $s = 3$ on input window sizes of $k \in \{6, 10\}$ and compares them to results of approaches with a similar sampling window size without striding. The decision to use a stride of 3 was made empirically. The results indicate that the performance is almost identical to the non-strided experiments and therefore it is possible to achieve similar performing results on a reduced frame rate. By using a stride $s = 3$ the effective

**Table 3.3**: Two strided approaches with an offset $s = 3$ and *input window size* of $k \in \{6, 10\}$ are compared with non-strided that share a similar sampling window size $\kappa$. The results show that the sampling window size is more important than the number of actual considered frames, since the performances does not decrease significantly when taking less frames for comparable $\kappa$.

| | $\kappa$ | kick | punch | wave | mAP |
|---|---|---|---|---|---|
| PBAR-MF18 | 18 | **0.33** | 0.40 | 0.64 | **0.46** |
| PBAR-MF6s3 | 16 | 0.32 | 0.44 | 0.54 | 0.43 |
| PBAR-MF24 | 24 | 0.30 | **0.45** | 0.59 | 0.45 |
| PBAR-MF10s3 | 28 | 0.28 | 0.41 | **0.65** | 0.45 |

frame rate is approximately one third of the original and hence around 10 frames per second.

## 3.4 Multi-class vs. binary problem formulation

For the evaluation of the binary problem formulation PBAR-MF10 was again chosen as baseline. As explained in Section 2.4.2, the number of classes was reduced. The idea behind this decision was that it might be easier for the network to distinguish between ordinary and non-ordinary activities. Splitting the available data in such manner would lead to more samples per class and hopefully a better performance. As Table 3.4 indicates this is the case. By

**Table 3.4**: For this experiment again the best performing architecture, PBAR-MF10, was chosen and trained in binary and multi-class manner. On the presented test set the binary approach performs approximately 18% mAP better than its multi-class counterpart.

| | mAP |
|---|---|
| multi-class | 0.48 |
| binary | **0.66** |

tackling the problem using a binary problem formulation the mean average precision could be increased on the regarded test set by about 18% mAP.

However, a major drawback of the binary formulation is that it loses the ability to distinguish between the kind of activity that was perceived. This makes it more difficult to understand the decision of the model, especially when the target class consists of a variety of different activities.

# 4 Conclusion

This report presents our work on Part Affinity Field based Activity Recognition. It gives an overview over how to include the information of the Part Affinity Fields provided by the OpenPose framework into a lightweight approach, which is designed to work for real-time applications. Furthermore, various topics like i) the impact of the number of input frames, ii) the impact of different simple dimensionality reduction approaches, and iii) a comparison between multi-class and binary problem formulation and how they influence the performance were evaluated. Future work will address further aspects in order to improve the performance and take a closer look on the temporal aspect of the approach: Does the usage of tracking algorithms improve the performance compared to the temporal cuboid approach? Can we benefit from the incorporation of Spatio-Temporal Affinity Fields [9]? How does MobileNet with 3D convolutions [6] perform? In addition to that, more elaborate yet fast dimensionality reduction approaches like PCA or LDA as well as incorporating an understanding of similarity of activities into the approach will be subject to future investigations.

# References

[1] Zhe Cao et al. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *CoRR* abs/1611.08050 (2016). arXiv: 1611.08050. URL: http://arxiv.org/abs/1611.08050.

[2] Barbara Hilsenbeck et al. “Action Recognition in the Longwave Infrared and the Visible Spectrum Using Hough Forests”. In: *IEEE International Symposium on Multimedia, ISM 2016, San Jose, CA, USA, December 11-13, 2016*. 2016, pp. 329–332. DOI: 10.1109/ISM.2016.0072. URL: https://doi.org/10.1109/ISM.2016.0072.

[3] Andrew G. Howard et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *CoRR* abs/1704.04861 (2017). arXiv: 1704.04861. URL: http://arxiv.org/abs/1704.04861.

[4] Vicky Kalogeiton et al. “Action Tubelet Detector for Spatio-Temporal Action Localization”. In: *ICCV 2017 - IEEE International Conference on Computer Vision*. Venice, Italy, Oct. 2017.

[5] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015. URL: http://arxiv.org/abs/1412.6980.

[6] Okan Köpüklü et al. “Resource Efficient 3D Convolutional Neural Networks”. In: *CoRR* abs/1904.02422 (2019). arXiv: 1904.02422. URL: http://arxiv.org/abs/1904.02422.

[7] H. W. Kuhn and Bryn Yaw. “The Hungarian method for the assignment problem”. In: *Naval Res. Logist. Quart* (1955), pp. 83–97.

[8] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324. URL: https://doi.org/10.1109/ICCV.2017.324.

[9] Yaadhav Raaj et al. “Efficient Online Multi-Person 2D Pose Tracking with Recurrent Spatio-Temporal Affinity Fields”. In: *CoRR* abs/1811.11975 (2018). arXiv: 1811.11975. URL: http://arxiv.org/abs/1811.11975.

[10] M. S. Ryoo and J. K. Aggarwal. *UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)*. 2010.

[11] Daniel Weinland, Rémi Ronfard, and Edmond Boyer. “Free viewpoint action recognition using motion history volumes”. In: *Computer Vision and Image Understanding* 104.2-3 (2006), pp. 249–257. DOI: 10.1016/j.cviu.2006.07.013. URL: https://doi.org/10.1016/j.cviu.2006.07.013.

# A Theoretical Model for Measuring and Sensor Characterization in Optical Spectroscopy

*Julius Krause*

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

Technical Report IES-2019-11

## Abstract

The optical and digital resolution, as well as the signal-to-noise ratio are important characteristics of optical spectrometers and available in data sheets. But how can an optical spectrometer system be selected for a specific application? The article shall serve as an aid to characterize optical spectrometers and hyperspectral cameras by introducing a benchmark calculation which indicates the measurement uncertainty of absorption bands.

## 1 Introduction

In optical spectroscopy, the wavelength depended intensity of light is measured. Due to the interaction between light and matter, the direction of the light propagation can change by elastic scattering processes. Furthermore, light can be absorbed by interaction with molecules, which changes the intensity of the light. The wavelength dependent probability of light scattering and absorption depends on the material properties of the sample. Therefore, it is possible

to determine material properties of the sample by recording its reflected or transmitted optical spectrum. Applications can be found in various fields like smart agriculture, food industry as well as in petro chemistry [9].

Due to the continuously advancing development of microsystems technology (MEMS), miniaturized spectrometers and hyperspectral camera systems can be manufactured cost-effectively and in large quantities. In order to achieve a comparability of sensors of different types, a benchmark parameter is presented below, which links the sensor noise with the optical and digital resolution.

In the following chapter the state of the art in chemometrics is briefly explained. Afterwards, signal generation and detection are discussed in more detail. Finally, the findings are used to define spectral features and sensor characterization.

# 2 State of the art in chemometrics

For the statistical analysis of spectroscopic data, the research discipline of chemometrics has developed within the field of chemistry. In the following, the state of research on theoretical simulation and in addition, the established pre-processing methods of chemometrics are referred.

Mainly core statements are given. For detailed information meaningful sources are given in each section.

## 2.1 Theoretical spectroscopy and simulation of spectroscopic results

Molecular vibrations can be excited by interaction with light, which causes an absorption of the light due to the law of energy conservation. For better understanding it is useful to consider light as particles, which are called photons. The energy of a photon is given by its frequency, which can also be expressed by a wavelength using the speed of light. And as result form quantum mechanics, only discrete energy levels of molecular vibrations can be excited. Both, the fundamental law of energy conservation and the discrete energy levels

of molecular vibrations lead to the simple result, that only photons with a wavelength, that matches these energy levels can be absorbed.

However, the anharmonic potential of atomic forces lead to an non-linearity in the energy levels of molecular vibrations. Therefore, the energy levels change strong in a solid-state or liquid sample caused by the presence of additional atoms, temperature or pressure. For this reason, theoretical spectroscopy is still a field of research. Simulation of spectroscopic results is only possible in case of simple molecules in solutions with a sparse concentration [1].

Furthermore, the transfer of chemometric calibration models to other products is quite impossible. This means, the calibration of sugar content of apples only can be applied to apples and not to other types of fruit.

## 2.2 Chemometric methods for spectral preprocessing

In the previous sections, the focus was on absorption and its relationship to material properties. However, the absorption can only be detected indirectly, whereas the reflected or transmitted light can be detected directly. Therefore, several methods have been developed to correct non-linearity of absorption, scattering effects and transfer of chemometric calibration models.

### 2.2.1 Absorbance units

In chemometrics, light which is not detected by the sensor $(1 - \boldsymbol{r})$ is referred as absorption, often this signal is also expressed in

$$\boldsymbol{a} := \log(1 - \boldsymbol{r}) \tag{2.1}$$

absorbance units (AU). Where $\boldsymbol{r}$ describes the reflected signal detected and discretized by the sensor and logarithms are used due to the exponential relationship between absorption and substance concentration by the Beer-Lambert-Law.

### 2.2.2 Scatter correction

In chemometrics, no distinction is made between the physical processes of elastic and inelastic light scattering. Only the terms absorption or reflection/transmission are used. Nevertheless, it is known that elastic scattering effects from Mie or Rayleigh theory have an impact on the spectrum and a scatter correction is necessary. Therefore, a Multiplicative-Scatter-Correction (MSC) or a Standard-Normal-Variate (SNV) is often applied as a pre-processing method [10]. Another approach is to derive the spectrum, which is often combined with smoothing operations [8, 4, 7].

### 2.2.3 Instrument transfer

An optical spectrometer records the spectrum of the light and converts it into a digital measurement signal. Depending on the instrument used, the spectrometers differ in their spectral range as well in their optical and digital sampling resolution. However, devices of the same type and manufacturer often differ in mechanical tolerances. For this reason, various methods for the transfer of calibration models have been developed [6, 3].

# 3 The spectral signature of a sample

The following section will describe the signal components of the optical spectrum in the near and short wave infrared $(780\,\text{nm} - 2500\,\text{nm})$. In the optical spectrum the physical effects of scattering and absorption are superimposed. Nevertheless, the spectrum can be evaluated by chemometric calibration models or machine learning methods. The amount of training data required for this can be reduced by making specific pre-assumptions. With the following model some physically motivated assumptions about properties (baseline, absorption bands) of the spectral signature (see fig.3.1) can be formulated.

This information model is used in chapter 5.1 to define characteristics. Finally, in chapter 5.2 a characterization of spectral sensor systems based on the detection of these features is proposed.

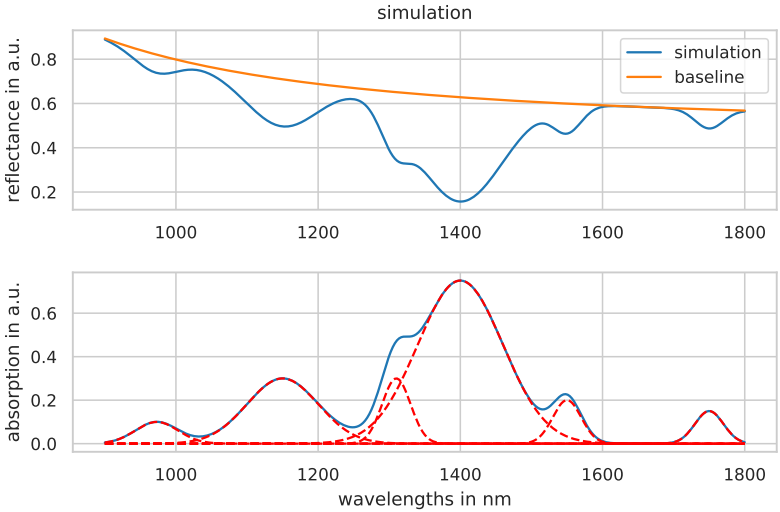



**Figure 3.1**: The spectral signature of an object is generated from the superposition of elastic and inelastic scattering (absorption). The absorbing molecule groups create Gaussian-shaped absorption bands. Their size allows quantitative analysis of the ingredients. Due to the wavelength-dependent elastic scattering processes (Mie- and Rayleight scattering), a smooth baseline is created.

## 3.1 A stochastic model to describe the spectral signature of a sample

The interaction between light and the sample can be described by a model of stochastic processes [5]. Therefore, the spectral signature of the sample is given by the probability density functions of $r_{\theta,\phi}(\lambda) \in [0,1]$ and transmission $t_{\theta,\phi}(\lambda) \in [0,1]$, depending on the wavelength $\lambda$, the angle $\phi \in [-\pi/2, \pi/2]$ of the incident light from the light source and the angle $\theta \in [-\pi/2, \pi/2]$ of the reflected or transmitted light. Both angles are related to the surface normal. A graph is used to describe the light and matter interaction (see fig.3.2): The light source radiates photons with the probability of $N_\phi(\lambda) \in [0,1]$ within the time period $T$ onto the sample. Multiple elastic scattering processes $s_{i,j}(\lambda) \in [0,1]$

can occur within and between different layers $i, j \in \mathbb{N}$ of the sample. For homogeneous materials without packaging, the number of layers can be reduced to one. From the surface, photons can be emitted in different angels $\theta$ as observable reflection and transmission. In addition, photons in each layer $n \in \mathbb{N}$ can be absorbed $a_n(\lambda) \in [0, 1]$ . For better readability the wavelength dependence is not explicitly referred at every point. The angles are usually unknown and cannot be measured. The angle-dependent scattering effects mainly appear when using very different samples or when comparing different measuring instruments. Therefore these quantities are given as an index.

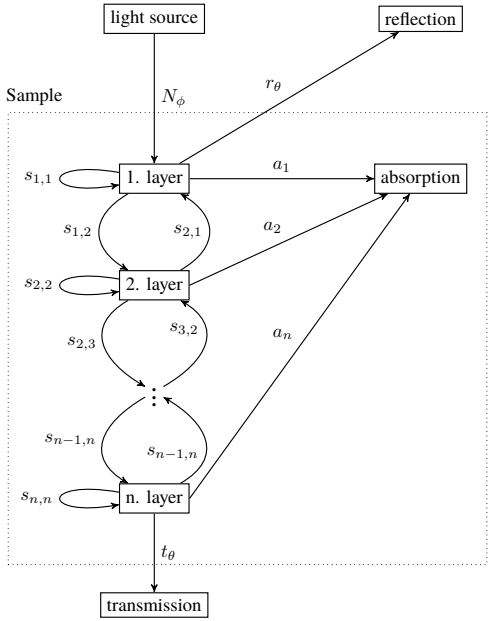



**Figure 3.2**: The spectral signature of the transmitted $t_\theta$ or reflected $r_\theta$ light from a sample is formed by multiple scattering $s_{i,j}$ and absorption $a_n$ processes within the sample. The scattering or absorption can differ in the different layers such as packaging, peel, pulp. In addition, the spectral response of the light source is also described as a probability density function $N_\phi$. The angles of incidence of the light source are named by $\phi$. The angles of emission of transmission and reflection are named by $\theta$.

The probability for emissions $a_i \in [0,1]$ in the state of absorption is depending on the concentrations $c_j \in [0,1]$ of absorbing molecules. So the chemical information is not directly observable. But energy conservation can be assumed, such that

$$\sum_n a_n + \int r_{\theta,\phi} d\theta + \int t_{\theta,\phi} d\theta = N_\phi \tag{3.1}$$

is valid. A general case for multiple light sources or directions from diffuse illumination can be created by adding a sum or integral over $\phi$.

Using eq. 3.1, some fundamental cases can be named:

- **Specular reflection** $\int r_\theta d\theta = N_\phi$: In case of specular reflection, all light is reflected. There is no transmission or absorption of the light.
- **Total absorption** $\sum_n a_n = N_\phi$: There is no measurement signal in case of total absorption.
- **Diffuse reflection** $\int t_\theta d\theta = 0$: This assumption is valid for samples of an infinite thickness. The reflected light is given by

$$\int r_\theta d\theta = N_\phi - \sum_n a_n$$

- **Diffuse transmission** $\int r_\theta d\theta = 0$: This assumption is valid for liquid samples. The transmitted light is given by $\int t_\theta d\theta = N_\phi - \sum_n a_n$

To minimize the angular dependency of the reflected signal, a diffuse illumination is usually used.

## 3.2 Absorption

The origin of absorption bands in the near and short-wave infrared are molecule groups with an polar hydrogen bonds like (OH, CH, NH, SH, COOH, ...) absorb the light. An absorption process becomes possible when the wavelength (energy) of the light matches the energy levels of the polar hydrogen bond within these funcional molecule groups.

The absorption

$$a_n(\lambda) = \sum_j^N \frac{c_j}{\sqrt{2\pi}\rho_j} \cdot e^{-\left(\frac{\lambda-\lambda_j}{\sqrt{2}\rho_j}\right)^2}, \tag{3.2}$$

is a sum over all absorbing molecule groups [2]. The energy levels $\lambda_i$ of the molecule groups mentioned are overlapping and also shifting non-linear depending on the sample composition. The with of the absorption band is given by $\rho_j$ and the concentration follows the Beer-Lambert law

$$c_j = 1 - e^{-\alpha_j} \tag{3.3}$$

with an absorption coefficient $\alpha_j \in \mathbb{R}$ depending on the dipole moment of the molecule. As described in the model (see fig.3.2), the absorption can also change in different layers, e.g. apple peel and fruit flesh. Therefore different absorption functions $a_i(\lambda)$ must be used.

However, the analysis of spectral data results in an ill-posed inverse problem: based on an detected absorption band, it is usually not possible to know which molecular group is the origin of the absorption.

## 3.3 Diffuse reflection and transmission

The scattering parameters $s_{i,j}(\lambda)$ of a sample vary depending on the microstructure (surface roughness and particle as well as molecule size). Using this scattering parameter, the reflected (transmitted) spectral signature

$$r_\theta(\lambda) = \left(1 - \sum_i a_i(\lambda)\right) s_{1,\theta}(\lambda) \tag{3.4}$$

results from light, which is not absorbed and scattered out of the top (bottom) layer of the sample. The scattering parameter can be explained by the Mie and Rayleigh theory. Because the required parameters such as illumination angle and measuring distance are not known in many cases, the scattering parameter s is assumed to be a continuous and smooth function. Furthermore, it is assumed that the scattering parameter in the region of an absorption band can be assumed to be locally constant.

# 4 Model for measurement systems in optical spectroscopy

The individual steps of signal generation are shown step by step in fig. 4.1. After the explanation of the spectral signature in the previous section, the spectroscopic measurement system is now in focus. An optical system is used to project the spectrum onto a detector. The detector converts the optical signal into a digital measurement signal.

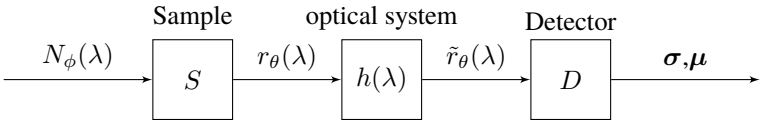



**Figure 4.1**: The $N_\phi(\lambda)$ photons emitted by a light source are reflected after interaction with the sample $S$. The angular and wavelength dependent reflectance of the sample forms the spectral signature $r_\theta(\lambda)$. An optical system (e.g. poly- or monochromator) is used to project the transformed reflectance spectrum $\tilde{r}_\theta(\lambda)$ onto a detector $D$.

## 4.1 Optical system

The optical resolution is diffraction-limited in the case of grating spectrometers and can be calculated with known grating, slit and distances. For this the Rayleigh criterion is used, the resolution limit $\Delta\lambda$ describes the radius of the Airy disk. However, this profile can also be well approximated by a Gaussian curve. Therefore, a spectral band $i \in \mathbb{N}$ of the optical system can be approximated with a *point spread function* (PSF) based on a Gaussian function

$$h_i(\lambda) = \frac{1}{\sqrt{2\pi}\rho_{\text{PSF}}} \cdot e^{-\left(\frac{\lambda-\lambda_i}{\sqrt{2}\rho_{\text{PSF}}}\right)^2} \tag{4.1}$$

which is mathematically easy to handle. In the case $\Delta\lambda = 0$ of an ideal optical system the transfer function $h_i(\lambda) = \delta(\lambda - \lambda_i)$ is generated. In data sheets the resolution of the optical system is usually specified by the *FWHM* (Full width (at) half maximum). Which is also related to $\rho_{\text{PSF}} = \frac{FWHM}{2\sqrt{2\ln(2)}}$.

The reflection signal

$$\tilde{r}_\theta(\lambda) = h(\lambda) * r_\theta(\lambda) =: \mu_p(\lambda) \tag{4.2}$$

which is projected onto the detector defines also the photon current which used in the next chapter.

Another property of the point spread function is the smoothing of the reflection signal. This leads to an attenuation of the absorption bands (eq. 3.2). Using eq. 4.2 and assuming Gaussian functions in eq. 4.1 and eq. 3.2 a new attenuated parameter for the concentration

$$\tilde{c_j} = \frac{\rho_j}{\sqrt{\rho_{\text{PSF}}^2 + \rho_j^2}} c_j \tag{4.3}$$

can be specified.

## 4.2 Detector model based on EMVA1288

The EMVA1288 standard contains a comprehensive description of the various signal contributions in semiconductor detectors and the digitization that follows. However, the EMVA1288 standard is used to characterize camera sensors without optics and refers to illumination with monochromatic light.

The noise (variance) of the grey values of a spectral band

$$\sigma_i^2 = K^2\sigma_d^2 + \sigma_q^2 + K\left(\mu_i - \mu_{i,\text{dark}}\right) \tag{4.4}$$

results from the amplified dark noise $\sigma_d$, the quantization noise 1/12 DN. The fluctuations of the photon stream are subject to a Poisson distribution and are signal dependent.

The signal

$$\mu_i = \int_{-\infty}^{\infty} \text{rect}\left(\frac{\lambda - \lambda_i}{\Delta\lambda}\right) \tilde{r}_\theta(\lambda)\eta(\lambda)Kd\lambda + K\mu_{\text{dark}} \tag{4.5}$$

of a spectral band results from the signal sampled over the range $\Delta\lambda$.

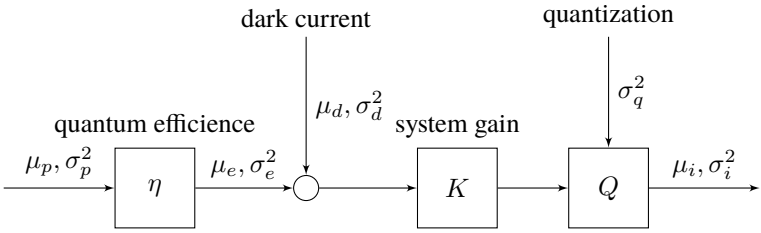



**Figure 4.2**: EMVA1288 Sensor system: The photon current $\mu_p$ is subject to a fluctuation $\sigma_p$. In the semiconductor, the photons generate electrical charge carriers $\mu_e$ with the quantum coefficient $\eta$. Thermal excitation produces an additional dark current $\mu_d$. The charge carriers are amplified analogously by the factor $K$ and then converted into a quantized measurement signal $\mu_i$ by an analog-digital-converter (ADC).

## 4.3 White and black balance

A white and black balance necessary is because of the spectral characteristics of the light source $N_\phi(\lambda)$, the quantum efficiency $\eta(\lambda)$, as well as the additional dark current of the detector $\mu_{y,\text{dark}}$. The signal

$$g_i = \frac{\mu_i - \mu_{i,\text{dark}}}{\mu_{i,\text{ref}} - \mu_{i,\text{dark}}} \tag{4.6}$$

can be calculated based on a reference spectrum of a sample with known reflectance and the dark signal. This wavelength dependent scaling leads to an amplification

$$\sigma_{g,i} = \frac{\sigma_i}{\mu_{i,\text{ref}} - \mu_{i,\text{dark}}} \tag{4.7}$$

of the noise of spectral bands. In many cases, a significant increase in noise can be observed at the borders of the spectrum.

# 5 Sensor characterization

In order to characterize a spectrometer system (see fig.4.1) the already introduced properties of light source $N_\phi(\lambda)$, optical system $h(\lambda)$ and detector $D$ will be combined in the following. The aim is an estimation of the measurement uncertainty for the detection of absorption bands.

## 5.1 Features in optical spectroscopy

The intensity of an absorption band can be used to quantify sample properties, as introduced above by Beer-Lambert's Law. Therefore, the absorption bands will be defined as features

$$m_j := \int_{-\infty}^{\infty} \frac{c_j}{\sqrt{2\pi}\rho_j} \cdot e^{-\left(\frac{\lambda-\lambda_j}{\sqrt{2}\rho_j}\right)^2} s_{1,\theta}(\lambda_j) d\lambda = c_j \cdot s_{1,\theta}(\lambda_j), \tag{5.1}$$

where the scattering parameter $s_{1,\theta}(\lambda_j)$ is assumed to be locally constant. These features are attenuated by the optical system and are recorded with noise. Using the relation $m_j \propto c_j$ and eq. 4.3 and 4.7 lead to a standard deviation

$$\sigma_{m_j} \propto \frac{\sqrt{\rho_{\text{PSF}}^2 + \rho_j^2}}{\rho_j} \frac{\sigma_g}{\sqrt{n}} \tag{5.2}$$

in the detection of spectral absorption bands. The optical attenuation of the absorption bands in the first term has an amplifying effect. Depending on the digital resolution, the noise influence is reduced by acquisition with $n$ channels.

## 5.2 Example for a new benchmark calculation in optical spectroscopy

From laboratory tests it is known that for recording moisture the feature $m$ at $\lambda_m = 1350\,\text{nm}$ with a width of $\rho_m = 50\,\text{nm}$ must be used. Two spectrometer systems with different characteristics are available. One system with low noise (sensor A) and high resolution (sensor B).

**Table 5.1**: Sensor comparison: Sensor A has a lower optical resolution, the spectral range of $900 - 1650$ nm is recorded with 128 bands. Sensor B has a high optical resolution, the spectral range of $900 - 1700$ nm is recorded with 255 bands. Due to the lower light per spectral band the noise of sensor B is increased compared to sensor A.

| | Sensor A | Sensor B |
|---|---|---|
| $\rho_{\text{PSF}}$ in nm | 20 | 5 |
| Bands/nm | 0.18 | 0.32 |
| $\sigma_g @ \lambda_m$ in nm | 0.01 | 0.02 |

By multiplying the digital resolution Bands/nm by the width of the absorption band $\rho_m$ the number of $n$ spectral bands involved in the sensor can be determined. This results in an estimated standard deviation of the feature $m$ with $\sigma_{m,\text{A}} = 0.0039$ for sensor A and $\sigma_{m,\text{B}} = 0.005$ for sensor B. For a general comparison of the two sensors the trend from $\sigma_m$ over $\rho_m$ is shown in figure 5.1.

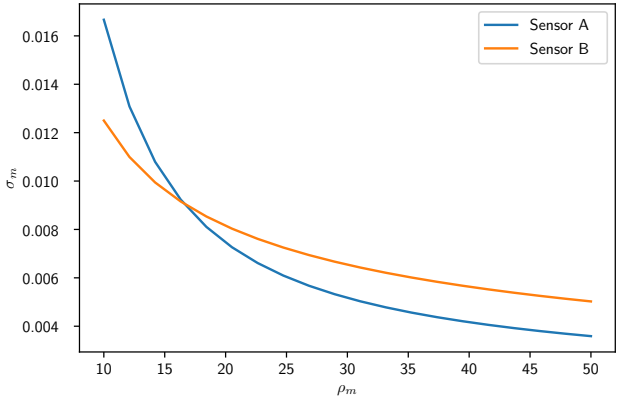



**Figure 5.1**: With increasing width of the absorption band which is to be detected, the number of spectral channels in the sensor increases, whereby the influence of the optical resolution also decreases in proportion. With a defined absorption band, the measurement uncertainty of the sensors can therefore be compared in the graph.

# 6 Summary

The signal generation in optical spectroscopy was described in terms of stochastic processes starting from the light source up to interpretable features. The focus was on signals in the near and short wave infrared spectrum. For quantitative statements on sample properties, the characteristics of absorption bands were justified and their signal portion was presented in reflection and transmission measurements.

Based on the absorption bands as quantitative features in optical spectra an estimation of the stochastic measurement uncertainty was formulated. For this purpose, the optical resolution was combined with the detector properties according to EMVA1288. As a result, spectroscopic measurement systems can be characterized by the expected stochastic measurement uncertainty. The definition of task-specific requirements for the resolution of certain absorption bands enables a benchmark for spectroscopic measurement systems as a whole. The approach can be generally used for hyperspectral cameras including illumination and optics or novel compact spectrometers from the consumer sector.

# References

[1] Krzysztof B. Be and Christian W. Huck. “Breakthrough potential in near-infrared spectroscopy: Spectra simulation. A review of recent developments”. In: *Front. Chem.* 7.FEB (2019), pp. 1–22. ISSN: 22962646. DOI: 10.3389/fchem.2019.00048.

[2] Adrian Jon Brown. “Spectral curve fitting for automatic hyperspectral data analysis”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.6 (2006), pp. 1601–1607. ISSN: 1558-0644. DOI: 10.1109/TGRS.2006.870435.

[3] Robert N. Feudale et al. “Transfer of multivariate calibration models: a review”. In: *Chemom. Intell. Lab. Syst.* 64.2 (2002), pp. 181–12. ISSN: 01697439. DOI: 10.1016/S0169-7439(02)00085-0.

[4] David W. Hopkins. “Revisiting the Norris Derivative Quotient Math in Regression”. In: *NIR news* 27.7 (2016), pp. 23–28. ISSN: 0960-3360. DOI: 10.1255/nirn.1643.

[5] Stéphane Jacquemoud and Susan Ustin. *Leaf Optical Properties*. Cambridge University Press, 2019. DOI: 10.1017/9781108686457.

[6] Jr Jerome J. Workman. “A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy”. In: *Applied Spectroscopy* 72.3 (2018), pp. 340–365. DOI: 10.1177/0003702817736064.

[7] Karl H. Norris. “Understanding and Correcting the Factors Which Affect Diffuse Transmittance Spectra”. In: *NIR news* 12.3 (June 2001), pp. 6–9. ISSN: 0960-3360. DOI: 10.1255/nirn.613.

[8] P. C Norris K. H; Williams. “Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size”. In: *Cereal chemistry* (1984). ISSN: 0009-0352.

[9] Yukihiro Okazaki. “Near-Infrared SpectroscopyIts Versatility in Analytical”. In: *Anal. Chem* 28.June (2012), pp. 545–562.

[10] Åsmund Rinnan, Frans van den Berg, and Søren Balling Engelsen. “Review of the most common pre-processing techniques for near-infrared spectra”. In: *TrAC Trends Anal. Chem.* 28.10 (Nov. 2009), pp. 1201–1222. ISSN: 0165-9936. DOI: 10.1016/J.TRAC.2009.07.007.

# High-NA Confocal Measurement by Diffractive Optical Elements

*Zheng Li*

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

Technical Report IES-2019-11

## Abstract

Diffractive optical elements (DOEs) can produce high-numerical-aperture (NA) spots over a large field. They can be combined with a low-NA objective to measure a large area with high resolution. This work shows experiments of using DOEs in reflection confocal microscope to resolve small structures beyond the capability of the objective. Both qualitative and quantitative results have shown enhancement in lateral and axial resolution by the application of the DOEs, which also agrees to the imaging theory of confocal microscopes.

## 1 Introduction

Confocal microscopy has been widely used as a standard measurement method in many fields for years [11]. The resolution of a confocal microscope is mainly dependent on the numerical aperture (NA) of the objective. Objectives with higher NAs can produce smaller illumination spots, and thus they can image the sample with high resolution. However, high-NA objectives are very limited

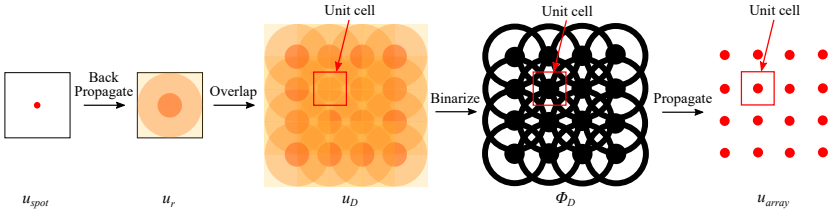



**Figure 1.1**: Design procedures of the DOE which generates spot array with overlapping apertures [4].

in the field of view and only a small portion of the sample can be measured at a time. Besides, high-NA objectives are also expensive due to complicated optical design and manufacture.

In order to solve the problems, diffractive optical elements are proposed to be used in confocal microscopes [3, 7, 6]. They can produce spots with overlapping apertures. Unlike the microlens array, the NA for a produced single spot is no longer limited by the unit area above it. In contrary, the surrounding area also contributes to the spot. So the proposed DOEs can produce a dense spot array with a high NA. The design procedures if the DOEs are described by Fig. 1.1 [5]. First, a required target spot field distribution $u_{spot}$ is defined. By simulation, the target field propagates back and forms a spherical-wave-like field distribution $u_r$. Rayleigh-Sommerfeld integral [8, 9, 10] is used as the simulation method for the diffraction field propagation, which is shown as

$$u_r = \iint_{\Sigma} u_{spot}(\mathbf{r}') \frac{e^{-ik|\mathbf{r}-\mathbf{r}'|}z}{|\mathbf{r}-\mathbf{r}'|^2} dx'dy', \tag{1.1}$$

where $\Sigma$ denotes the surface on the boundary, i.e. the plane which $u_{spot}$ lies on and the semi-infinite sphere behind it, $\mathbf{r} = (x, y, z)$ is the coordinate of $u_r$, $\mathbf{r'} = (x', y', z')$ is the coordinate on $\Sigma$, and $k$ is the wave number.

By utilizing the idea of overlapping aperture, $u_r$ is duplicated and overlapped with a designed pitch to form the overlapping field $u_D$. Then the phase of $u_D$ is extracted and binarized with a binarization factor $B$ [3] as follows,

$$\phi_D(x, y) = \mod\left(\left\lfloor \frac{\arg[u_D(x, y)] + B}{\pi} \right\rfloor, 2\right)\pi. \tag{1.2}$$

Finally the binarized field propagates back by simulation to validate the design result. In this case, a dense spot array with a high NA can be generated by plane wave illumination. And the produced spots with an NA up to 0.77 has been demonstrated [5]. However, when being used in a confocal setup like Fig. 1.2, such DOEs introduce significant disturbances in the imaging path. In order to realize the setup, the DOEs need modification to increase the zero-order diffraction which allows the generated spots to be imaged through itself. This is done by adding a plane-wave component to the overlapped field $u_D$,

$$u'_D = u_D + W, \tag{1.3}$$

where $W$ is a constant which is optimized iteratively to achieve the best signal-to-noise ratio of the spots in the image. In this way, the disturbance added by the DOE is significantly reduced when the spots are imaged through it, which has already been demonstrated by experiments [5]. The resulting DOE, which is called See-through DOE, can thus be used in the confocal microscope in Fig. 1.2. And both lateral and axial resolution can be enhanced as shown by theory and experiments in the following chapters.

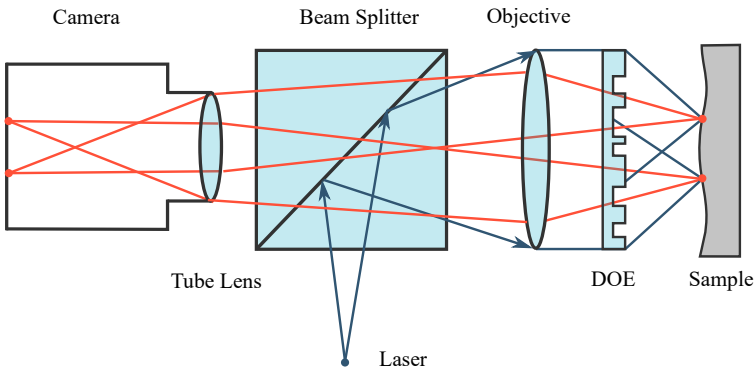



**Figure 1.2**: Setup of a confocal microscope using the DOE. A laser is collimated by the objective to illuminate the DOE. The produced spots are again imaged by the objective onto the camera sensor.

# 2 Resolution enhancement by the DOE

The confocal microscope setup in Fig. 1.2 uses a low-NA objective to increase the field of view and a see-through DOE to project high-NA spots. The objective also acts as a collimator to produce plane-wave illumination for the DOE. It is suitable for both opaque surface measurement and fluorescence microscopy, which transmission microscopes cannot measure. When an opaque surface is measured, lateral resolution can be significantly increased by the DOE while axial resolution can only be slightly increased [4]. When fluorescent samples in transparent medium are measured, both lateral and axial resolutions can be increased which is comparable to a high-NA objective.

## 2.1 Theory of scanning microscopy

The image formation of a scanning microscope can be described as the following equation[12]:

$$U(x_2, y_2; x_s, y_s) \tag{2.1}$$

$$= \iint_{-\infty}^{\infty} h_1(x_0, y_0) t(x_0 - x_s, y_0 - y_s) h_2\left(\frac{x_2}{M} - x_0, \frac{y_2}{M} - y_0\right) dx_0 dy_0, \tag{2.2}$$

where $(x_0, y_0)$ is the object coordinate, $(x_2, y_2)$ is the image coordinate, $(x_s, y_s)$ is the scanning position, $h_1(x_s, y_s)$ is the illumination point spread function (PSF), $h_2(x_s, y_s)$ is the imaging PSF, and $t(x_0, y_0)$ is the object transmissivity or reflectivity. For a confocal microscope, a point detector is used at $x_2 = y2 = 0$ and the intensity at every scanning position can be expressed by

$$I(x_s, y_s) = \left| \iint_{-\infty}^{\infty} h_1(x_0, y_0) t(x_0 - x_s, y_0 - y_s) h_2\left(-x_0, -y_0\right) dx_0 dy_0 \right|^2, \tag{2.3}$$

which can be simplified to the following equation because the PSF is even,

$$I(x_s, y_s) = \left| h_1 h_2 * t \right|^2, \tag{2.4}$$

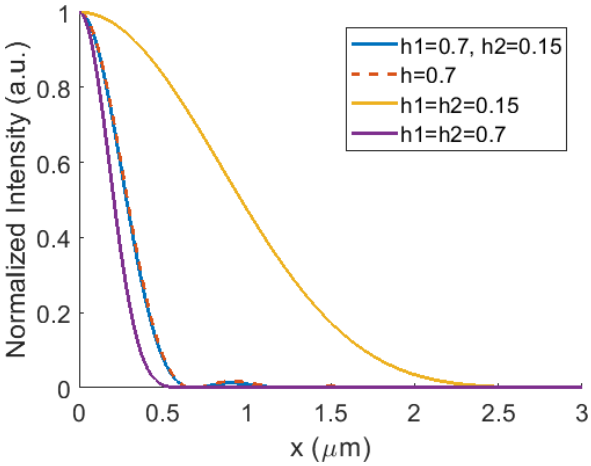
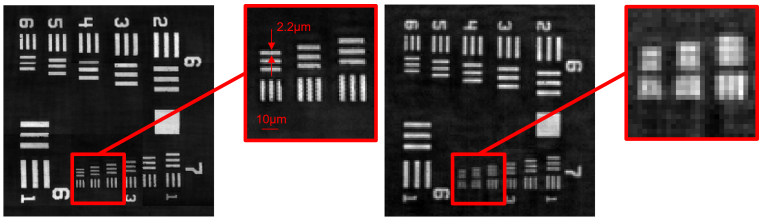



**Figure 2.1**: Simulation of a single point with different illumination and imaging PSFs. The dashed line represent a wide-field microscope configuration with a single PSF.

The equivalent PSF for the confocal image thus becomes $h_1 h_2$. In this case, if the illumination $h_1$ is high-NA and the imaging $h_2$ is low-NA, the combined PSF will still be dominated by the high-NA illumination.

Figure 2.1 shows the simulation results of the lateral intensity profile when a single point is imaged by illumination and imaging with different NAs. It is shown that when the illumination has high NA, the combined confocal PSF is independent of the low-NA imaging objective and is slightly smaller than the wide-field high-NA curve. Thus the lateral resolution can be increased by such a setup in Fig. 1.2. This is also very similar to the principle of super-resolution microscopy like STED [2] or PALM [1]. Similarly, the axial resolution can also be increased for a point-like object in fluorescence microscopy, which has been explained in [4].




(a) DOE confocal scanning image. (b) Wide-field image.

**Figure 2.2**: Images of a resolution target taken by an objective with NA=0.15.

## 2.2 Experiments of the DOE in confocal microscopy

Experiments are also made to test the resolution enhancement by the DOE in a confocal microscope as shown in Fig. 1.2. A standard positive USAF resolution target from Thorlabs is used as the test object with a maximum resolution of 228 line pairs per millimeter. The target is scanned laterally in a zig-zag way and a confocal scanning image is obtained.

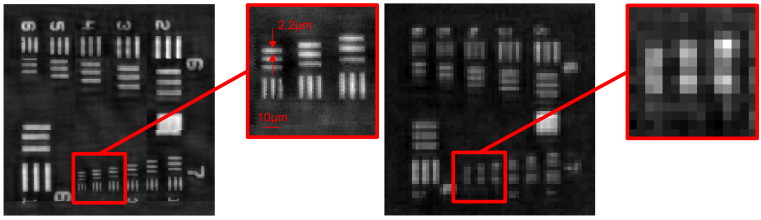
Fig. 2.2 shows a comparison of a DOE confocal scanning image and a wide-field image both taken by an objective with NA=0.15. It is obvious that the DOE increases the lateral resolution and the even finest patterns can be clearly resolved. The zig-zag like artifacts in the image are caused by insufficient accuracy of the xy stages which leads to the misalignment in the confocal image reconstruction. Contrarily, the wide-field image is totally blurred because the numerical aperture of the objective is not high enough.

Furthermore, images shown in Fig. 2.3 are also taken by an objective with an even smaller NA of 0.07. There is still a very obvious resolution enhancement. However, the signal-to-noise ratio and the resolution of the confocal scanning image are also slightly reduced compared to the image taken by an objective with an NA of 0.15.

There could be several reasons for this. First, the diffraction efficiency of a binary DOE is limited. There is unavoidable -1 order diffraction which is stray light and will be collected by the objective to form a noisy background. When

the NA of the objective is lower, the signal is weaker so the signal-to-noise ratio is reduced. This phenomenon can be mitigated by using a multi-level DOE to increase the diffraction efficiency. Moreover, the Andor Zyla 5.5 camera we used has a large pixel size of $6.5\,\mu\mathrm{m}$. And we use a $0.63\times$ tube lens and a $2.5\times$ objective with NA=0.07. The total magnification is 1.575, which is pretty small. This leads to a relatively large equivalent pinhole size of roughly $3.7\,\mu\mathrm{m}$, which cannot effectively block the stray light. By using a camera with a smaller pixel size or a tube lens with a larger magnification can mitigate the problem.

After the lateral measurement, the axial resolution is also tested with the fluorescence microscope setup shown in Fig. 2.4. The Sphero Rainbow fluorescent particles are used as samples. The excitation wavelength is $630\,\mathrm{nm}$ and the emission wavelength is from $672\,\mathrm{nm}$ to $712\,\mathrm{nm}$. The sizes of the beads are $3.0$-$3.4\,\mu\mathrm{m}$. In the experiment, only one fluorescence bead is focused. The bead is moved vertically to measure the intensity response. Both objectives with NA of 0.15 and 0.07 are used for testing. The produced spots have axial full width at half maximum (FWHM) of $19.5\,\mu\mathrm{m}$ and $17.2\,\mu\mathrm{m}$ respectively, which corresponds to a NA of roughly 0.25, because of the different collimation quality of the objectives. The confocal signals show axial FWHM of $25.9\,\mu\mathrm{m}$ and $26.3\,\mu\mathrm{m}$ respectively. The results show that the axial resolution is almost independent on the imaging objective, which is predicted by the theory. Still the confocal axial peaks are a bit wider than the illumination spot. The reason can be that the pixel as a pinhole is large, and the diameter of the bead is not negligible.




(a) DOE confocal scanning image. (b) Wide-field image.

**Figure 2.3**: Images of a resolution target taken by an objective with NA=0.07.

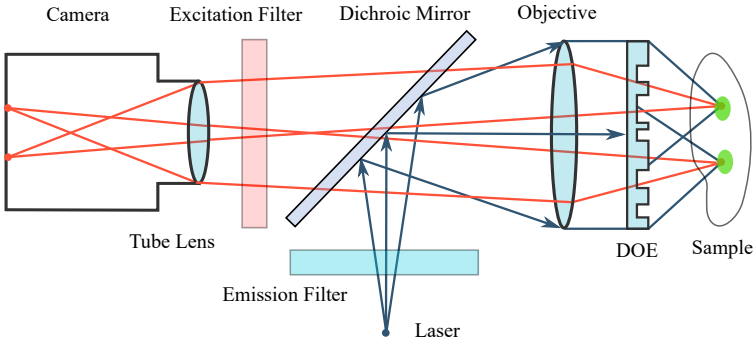
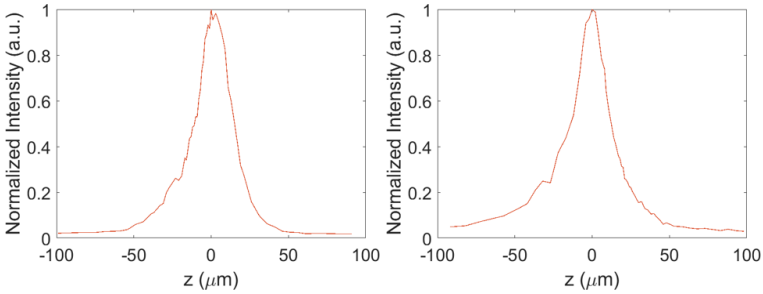



**Figure 2.4**: Setup of a fluorescence confocal microscope using the DOE.




(a) Axial response by an objective with NA=0.15.

(b) Axial response by an objective with NA=0.07.

**Figure 2.5**: Axial intensity response of an in-focus fluorescent bead.

# 3 Conclusion

Traditional confocal microscopy relies on high-NA objectives to achieve high resolution. However, high-NA lenses have a very limited field of view. The See-through DOE can be used with a low-NA objective in a reflection confocal microscope to provide a large field of view. The DOE can produce high-NA spots and maintain the resolution of a high-NA objective in such a setup.

For surface measurements, 2D scan was performed and the enhancement of resolution is clearly demonstrated. Meanwhile, the See-through DOE is also successfully used in fluorescence microscopy. Fluorescence signals of the beads were observed and also axial response was tested. The axial FWHMs are independent of the NA of the objective, which also agrees with the theory.

In the future, for surface measurement, the measurement uncertainty will be tested. For fluorescence measurement, a 3D scan of living cells will be carried out. New experiments are planned to further demonstrate the capability of the DOEs to increase the measurement resolution.

# References

[1] Eric Betzig et al. “Imaging intracellular fluorescent proteins at nanometer resolution”. In: *Science* 313.5793 (2006), pp. 1642–1645.

[2] Stefan W Hell and Jan Wichmann. “Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy”. In: *Optics letters* 19.11 (1994), pp. 780–782.

[3] Bas Hulsken, Dirk Vossen, and Sjoerd 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. 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. 25–46. ISBN: 978-3-7315-0936-3.

[5] Zheng Li et al. “Application of DOE in confocal microscopy for surface measurement”. In: *Photonics and Education in Measurement Science 2019*. Ed. by Maik Rosenberger, Paul-Gerald Dittrich, and Bernhard Zagar. Vol. 11144. International Society for Optics and Photonics. SPIE, 2019, pp. 254–261. DOI: 10.1117/12.2531610. URL: https://doi.org/10.1117/12.2531610.

[6] Xiyuan Liu and Karl-Heinz Brenner. “High Resolution Wavefront Measurement with Phase Retrieval Using a Diffractive Overlapping Micro Lens Array”. In: *Fringe 2013*. Springer, 2014, pp. 233–236.

[7] Xiyuan Liu, Tim Stenau, and Karl-Heinz Brenner. “Diffractive micro lens arrays with overlapping apertures”. In: *Information Optics (WIO), 2012 11th Euro-American Workshop on*. IEEE. 2012, pp. 1–2.

[8] Fabin Shen and Anbo Wang. “Fast-Fourier-transform based numerical integration method for the Rayleigh-Sommerfeld diffraction formula”. In: *Applied optics* 45.6 (2006), pp. 1102–1110.

[9] Arnold Sommerfeld. *Mathematical Theory of Diffraction*. Boston, MA: Birkhäuser Boston, 2004, pp. 9–68. ISBN: 978-0-8176-8196-8. DOI: 10.1007/978-0-8176-8196-8_2. URL: https://doi.org/10.1007/978-0-8176-8196-8_2.

[10] Arnold Sommerfeld. “Mathematische theorie der diffraction”. In: *Mathematische Annalen* 47.2 (1896), pp. 317–374.

[11] Jeroen Vangindertael et al. “Super-resolution mapping of glutamate receptors in C. elegans by confocal correlated PALM”. In: *Scientific reports* 5 (2015), p. 13532.

[12] Tony Wilson and Colin Sheppard. *Theory and practice of scanning optical microscopy*. Vol. 180. Academic Press London, 1984.

# A Realistic Predictor for Pedestrian Attribute Recognition

*Andreas Specker*

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

Technical Report IES-2019-11

## Abstract

The application of video surveillance systems in public areas to ensure public security is becoming increasingly important. A major task when evaluating the arising amount of video data is to find the occurrences of a person-of-interest on the basis of a testimony. For the comparison of a person's description with persons in the video data, the attributes of all persons must be recognized automatically. However, typical approaches to pedestrian attribute recognition simply predict all attributes for a person, regardless the visibility of relevant attributes. To address this problem, the concept of realistic predictors is used in this work to determine and improve the reliability of pedestrian attribute recognition.

## 1 Introduction

Nowadays, more and more video surveillance systems are used to ensure public security. Due to the large amount of image and video footage that is recorded by

(a) Detections (b) Viewpoints (c) Appearance (d) Occlusions

**Figure 1.1**: Different challenges in recognizing pedestrian attributes. Poor detections and occlusions can lead to only partially visible persons in images. Moreover, some attributes like backpack may not be visible from all point of views and attributes such as handbags may appear as many different types.

such systems, manual evaluation is hardly possible, which is why intelligent and automatic analysis systems are required. One of the most important evaluation tasks that can be automatically solved by applying convolutional neural networks (CNN) is person re-identification which aims to find all occurrences of a person-of-interest in the data. Typically, such a search is performed based on a cropped image of the person the system operators are interested in. But since it is not possible to cover all areas with CCTV cameras, one cannot be sure that a query image of the person-of-interest is always available. Thus, in such cases, descriptions of the semantic attributes are the only clues on which the person search can be based. The query attributes can be easily and directly extracted from witness descriptions. In order to find all persons corresponding to the obtained attributes, the semantic attributes of the persons present in the surveillance material must be recognized.

This pedestrian attribute recognition in an uncooperative, real-world scenario suffers from a lot of different challenges. Some of the most severe issues to overcome are visualized in Figure 1.1. Stable recognition of a person's semantic attributes is only possible if clean cutouts are available. But sometimes person detectors provide bad detections which show a lot of background clutter or only parts of a human body. Moreover, the view angle is an factor that greatly influences the appearance of a person. Attributes as for instance backpack may not be visible from every point of view. Similar issues arise from occlusions which make it difficult or impossible to determine certain attributes. Lastly,

attributes, such as handbag in Figure 1.1(c), can differ greatly regarding their appearance. Handbags come in different sizes and colors making the recognition task harder.

All those challenges indicate that meaningful attribute predictions can not be given in all cases. If, for instance, the lower-body of a person is occluded by a vehicle, no well-founded statement about the length of the lower-body clothing can be made. Although this is a very important topic, it is not present in existing literature regarding to pedestrian attribute recognition. However, with regard to typical one-hot classification, Wang et al. [17] present an approach which takes into account the hardness of the input images and only provides classification results if a reliable estimation is possible. Since attribute recognition, albeit multi-class, is a classification problem as well, the core idea of this work is to transfer and adapt the concept of realistic predictors to this task.

# 2 Related work

Generally, pedestrian attribute recognition approaches from related literature can be roughly divided into three different categories: global, part-based and attention-based methods.

**Global Models** Especially early deep learning-based works on pedestrian attribute recognition predict semantic attributes on solely a whole body image of a person. In [16] for instance, a multi-branch architecture is applied that contains a separate classification layer and loss for each attribute. In contrast, some works showed that it is advantageous not to learn all the attributes separately but instead learn them all together [7] or partitioned in groups of corresponding attributes [1]. In addition to that, the authors in [7] propose to weight the attributes during loss calculation according to their frequency of occurrence in the dataset to handle the large imbalances of attribute values. The results of newer works [15], however, indicate that with the development of larger CNN models the joint learning of attributes is not always beneficial and higher accuracies can be achieved if separate networks are used for different attributes. In general, global models are simple and therefore very efficient compared to more complex architectures. These results in faster training and testing, though

only using coarse information. Differences between global attributes, as gender, and small-scale attributes such as shoes or glasses are not taken into account and aggravate the recognition task.

**Attention-based Models** Attention-based methods aim to guide the network to focus on the most important regions of activation maps or features. [12, 13] propose networks that are capable of implicitly learning visual attention maps. A special feature of [12] is the use of a multi-directional attention mechanism which means that attention is shared between different semantic layers of the network. Moreover, Sarfraz et al. [14] introduce an approach to learn view-sensitive embeddings since the viewpoint of a person is really important with respect to the appearance of attributes. To improve attention maps explicitly, in [5] attention maps are refined using a exponential loss function. Although some attention-based methods are proposed in literature, the gain in accuracy is still limited compared to other research fields such as for instance person re-identification.

**Part-based Models** Part-based algorithms jointly leverage local and global information to improve recognition accuracy. This is done by either localizing body parts of persons using an extern [4, 9] or intern [3, 11, 18] module. In [4] patches obtained from a part detector are fed into a fine-grained classification model. Similar to that, [10] proposes to use the detector features of the whole person and detected parts as input patches for attribute classification layers. A slightly different way is followed in [9]. Instead of bounding boxes estimated by a body part detector, pose key points are exploited to localize meaningful body part regions. In contrast to these approaches, [18] introduces a method by which part localization and attribute classification is jointly learned in an end-to-end manner. In [3], the authors use mid-level image patches as representations of human body parts. Moreover, LGNet is presented in [11]. Consisting of a global and a local network branch, part detection is performed by creating so-called EdgeBoxes that are applied in a Region-of-Interest pooling module. Such part-based models are less efficient compared to simple global models but instead are able to focus on fine-grained information which is very important for recognition of very local attribute, as for instance glasses or shoes. However, it is important that body parts can be accurately detected because otherwise the approaches suffer from focusing on irrelevant regions of the input image.

Although part-based models implicitly handle the visibility of body parts or attributes, none of the approaches in literature deal with the fact that in a uncooperative real-world scenario attributes cannot be predicted for imperfect person image crops or occluded body parts. Therefore this work aims to close this research gap by investigating the concept of realistic predictors which is detailed in the following.

# 3 Methods

In this chapter, the baseline classification model is presented followed by a detailed description of the realistic predictor approach.

## 3.1 Baseline model

The baseline model is based on the typical classification pipeline for global pedestrian attribute recognition. Images are pre-processed and data augmentation is performed. Afterwards, images are fed into a backbone network with appended fully-connected classification layer and output probabilities are computed using the sigmoid function. In this case, the task is considered a multi-class classification task which means that all attributes are simultaneously predicted using a single classification layer. Sigmoid cross-entropy loss function (SCEL) is applied to train the CNN model. To handle the imbalanced distribution of positive attribute labels in the dataset, a weighting factor is added to the loss computation as proposed in [7]. Let $y_i^c \in [0, 1]$ be the target label of the $c$th attribute of the $i$th sample and $p^c$ the positive ratio of this attribute in the dataset. Then the weighting factor $w_i^c$ can be computed independently for each attribute and input image as follows:

$$w_i^c = \begin{cases} \exp(\frac{(1-p^c)}{\sigma^2}) & , \text{ if } y_i^c = 1 \\ \exp(\frac{p^c}{\sigma^2}) & , \text{ if } y_i^c = 0 \end{cases} \tag{3.1}$$

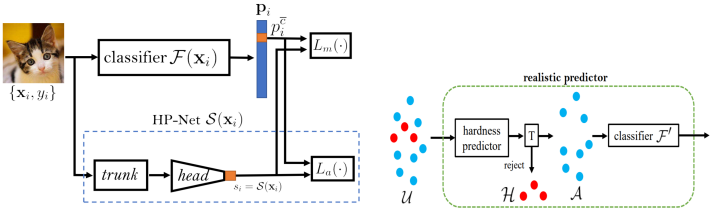



**Figure 3.1**: The general idea of realistic predictors. On the left, the architecture is shown consisting of two branches: a classification and a hardness prediction one respectively. The figure on the right depicts the testing stage. Samples with a hardness score above a threshold $T$ are discarded and not fed into the classifier. [17]

$\sigma$ stands for a hyperparameter which is set to 1 in all experiments. This weighting factor ensures that the network focuses on rare attributes by increasing the weight of such samples.

## 3.2 Realistic predictors

The concept of realistic predictors is adapted from [17]. The general approach is visualized in Figure 3.1. A network with two branches was designed to simultaneously train a classifier and a so-called hardness predictor. The classifier outputs probabilities $p_i$ for each class whereas the hardness prediction network computes hardness scores. Hardness scores $s_i$ are understood as predictions of the difficulty of the classification task for a specific input image. So, for instance, the hardness score should be higher if an object is only partially visible in comparison with a clean cut of the object of interest. The testing protocol is visualized in Figure 3.1 on the right. First, the hardness for all samples is predicted. To find those images for which no reliable classification can be provided, hard samples are discarded based on a threshold $T$. The remaining samples are then forwarded through the classifier and a class prediction is produced. In practice, only attributes for which the classifier is certain would be output and then used for person retrieval.

Two different losses are used to train the two network branches. For training the classifier, the use of a weighted softmax cross-entropy loss function is proposed. This loss function $L_m$ is shown in the following equation where $N$ stands for the number of samples in the batch and $p_i^{\bar{c}}$ depicts the predicted probability for target class $\bar{c}$ and sample $i$.

$$L_m = -\sum_{i=1}^{N} s_i \log p_i^{\bar{c}} \tag{3.2}$$

As mentioned earlier, the original paper deals with a one-hot classification problem in contrast to the pedestrian attribute task. Persons have several attributes at the same time, like a woman wearing a red shirt and blue jeans, and thus multiple classes can be true. Therefore the loss function for the multi-class task is adapted as follows:

$$L_m = -\sum_{i=1}^{N}\sum_{c=1}^{C} [y_i^c \log p_i^c + (1 - y_i^c) \log(1 - p_i^c)] \tag{3.3}$$

In addition to the sum over all samples, the sum of cross-entropy losses for all attributes is computed. $C$ denotes the number of different semantic attributes in this case and $y_i^c \in [0, 1]$ is the target label of the $c$th attribute.

Another alteration that was made is that the feedback of the predicted hardness score $s_i$ is omitted in contrast to the original paper. Whereas the authors propose this term to focus on those samples that are particularly hard during training, this is not necessarily beneficial for attribute recognition. In the object classification approach one can be certain that the object is actually present and visible in the input image. In contrast, especially small-scale attributes are often occluded and therefore not visible which could lead to a decrease in recognition accuracy if such samples are preferred during the training process. The network would not be able to base its decision on meaningful clues and to learn important information.

For training the hardness predictor, another loss function is proposed in [17].

$$L_a = -\sum_{i=1}^{N} \left[p_i^{\bar{c}} \log (1 - s_i) + \left(1 - p_i^{\bar{c}}\right) \log s_i\right] \tag{3.4}$$

The goal of this function is to produce large hardness scores if and only if the cross-entropy loss of the classification branch is high and vice versa. Therefore, a kind of inverse cross-entropy loss is used. The loss function gets minimal if $s_i = 1 - p_i^{\bar{c}}$ applies. In words, the hardness score is forced to be equal to the classification error measured by the prediction probability. Moreover, the more the estimated class probability differs from the target value the higher the loss of the hardness predictor.

Analogous to the classification loss function, the hardness predictor loss calculation has also be modified to match the requirements of the multi-class attribute classification problem. Again, the loss function is expanded to consider each attribute. Since in contrast to the one-class classification problem not only one positive class is relevant but instead the presence as well as the absence of all attributes, loss calculation is also based on the target label, as can be seen in the equation hereafter.

$$L_a = -\sum_{i=1}^{N}\sum_{c=1}^{C}\left[\Delta p_i^c \log\left(1 - s_i^c\right) + \left(1 - \Delta p_i^c\right)\log s_i^c\right], \tag{3.5}$$

$$with \Delta p_i^c = \left|y_i - p_i^c\right| \tag{3.6}$$

Thereby, the hardness predictor learns to estimate the difficulty of an image regardless of an attribute being present or not in the training image. This is ensured by applying the absolute value of the difference between the target class label $y_i^c$ and the predicted probability of the presence of an attribute $p_i^c$ instead of using $p_i^c$ directly.

Since the training of the hardness predictor network also suffers from data imbalances, DeepMAR weighting can be applied here as well, thus reducing the influence of unbalanced attributes distribution on the training.

## 3.3 Determination of thresholds

To improve the accuracy of pedestrian attribute recognition, meaningful thresholds for hardness scores need to be determined. It is important to find a good

trade-off between improving accuracy and rejecting as few samples as possible. Thus, multiple strategies to seek for meaningful thresholds are proposed and compared in the evaluation chapter. The thresholds are computed for each attribute independently making use of the evaluation data. To avoid that too much samples of an attribute are discarded, optimization is stopped as soon as the threshold is below that of the quantile rejection method.

**Threshold rejection** As a baseline for comparison of the other rejection approaches, one single threshold which is applied to all attributes is determined.

**Quantile rejection** In contrast, quantile rejection method sets the thresholds to a value so that a predefined portion of validation samples is discarded. Since the distribution of the hardness scores may vary between validation and testing data, the proportion of rejected samples can differ during testing stage.

**Mean accuracy / F1 rejection** This rejection approach aims to optimize the target evaluation metric, either mean accuracy or F1 score. The threshold value is lowered until the mean accuracy no longer increases or until the stop criterion mentioned above is reached.

# 4 Evaluation

The previously introduced approaches are evaluated and discussed in the following. After some details about the datasets used and the experimental setup, the results of the experiments are presented.

## 4.1 Datasets

The experiments are conducted on two different publicly available datasets. Both datasets contain person bounding boxes that are all taken from videos of surveillance cameras. A brief introduction to RAP-2.0 and PA-100K datasets is given in the following. Some sample images of both dataset can be found in Figure 4.1.

**Figure 4.1**: Randomly selected images from the datasets are shown for comparison. Figures (a) - (e) are taken from the RAP-2.0 dataset whereas Figures (f) - (j) are from the PA-100K dataset.

The **RAP-2.0** [8] dataset consists of 84,928 images taken from 26 different cameras. All cameras were mounted indoor and show scenes of a shopping mall. 72 different binary attributes ranging from gender to attachments are annotated. Since the distributions of the attribute annotations are highly unbalanced, only those attributes with a positive ratio greater than 1 % are used in the experiments. After discarding very rare attributes, 54 attributes remain whose positive ratios are shown in Figure 4.2.

Unlike the RAP-2.0 dataset, the **PA-100K** [12] dataset contains images recorded in an outdoor setting. According to the dataset name, 100,000 images from 598 different cameras are included and 26 binary attributes are provided. Moreover, distributions of attribute annotations are more balanced.

## 4.2 Experimental setup

**Data pre-processing and augmentation** During training phase, images are resized and randomly cropped to match the input size of the CNN. In addition, random flipping is applied to increase the diversity of training data.

**Backbone model** Experiments with different backbone models were carried out. Since the observations presented in this chapter are valid regardless of the CNN model used, only results for ResNet-50 [6] are presented and discussed.

**Parameters** To train the models, a multi-step scheduling scheme was applied in all experiments. Two steps are performed with a decay factor of 0.1. RAP-2.0

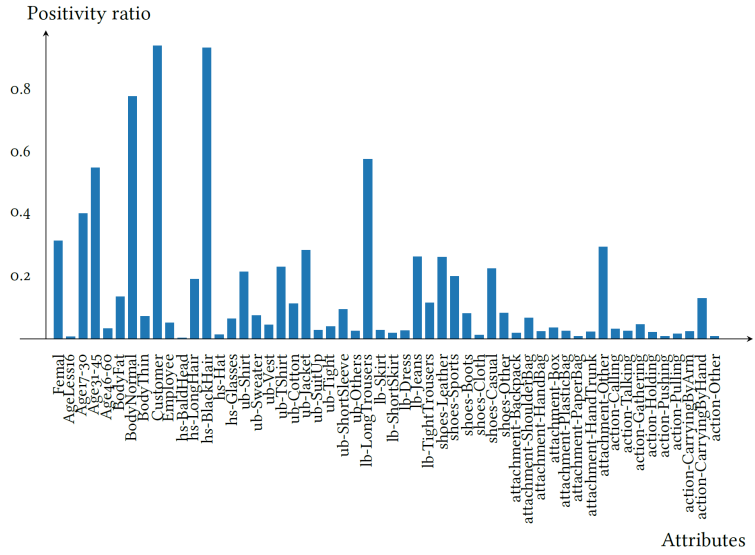



**Figure 4.2**: Positive ratios of RAP-2.0 attributes. Only few attributes have balanced distributions while most attributes such as *attachment-backpack* occur very rarely.

models were trained for a total of 180 epochs with steps after 60 and 120 epochs. The learning rate for the Adam optimizer was initially set to $10^{-4}$ for the classifier and $10^{-5}$ for the hardness predictor, respectively. For training the networks with the PA-100K dataset, parameters were set to the values suggested in [2].

## 4.3 Hardness prediction

Table 4.1 presents the attribute recognition results of the classifiers. Using positive ratio-based DeepMAR weighting of the loss during training significantly increases the recognition performance by reducing the influence of imbalanced attribute distributions. Moreover, the results clearly indicate that using feedback of the HP-Net for training the classifier network is not beneficial for pedestrian

**Table 4.1**: Quantitative evaluation of baseline methods on RAP-2.0 dataset. DeepMAR weighting of training samples greatly improves mA. Training the classifier with HP-Net feedback deteriorates the results in all metrics.

| Model | mA | Acc | Prec | Rec | F1 |
|---|---|---|---|---|---|
| SCEL | 64.29 | 62.26 | **82.55** | 70.09 | 75.81 |
| DeepMAR | **73.05** | **63.99** | 77.01 | **77.17** | **77.09** |
| SCEL + HP-Net feedback | 61.93 | 52.00 | 69.33 | 66.81 | 68.05 |
| DeepMAR + HP-Net feedback | 67.32 | 61.18 | 76.74 | 73.49 | 75.08 |

attribute recognition. In the original approach this feedback was proposed to force the classifier to focus on those samples which are hard to classify. But in contrast to typical image classification, attributes are small-scale features and thus not necessarily visible in hard-to-classify images. As a results, focusing on such hard samples confuses the CNN and accuracy decreases regarding all metrics as can be seen from the experimental results in the table.

Next, it is important to evaluate the quality of the given hardness predictor. For this purpose, Figures 4.3 and 4.4 show person images assessed as easy as well as hard are displayed. Figure 4.3 visualizes samples for the gender attribute. The qualitative results seem reasonable. It is easy for the classifier to classify a person as a woman if the person is wearing a skirt or has long hair that is clearly visible. In contrast, hard samples are images showing only partial persons such as the first image in Figure 4.3(b). Also a human cannot make a reliable statement about the sex, because only the legs of the person are visible. Moreover, images on which the length of the hair is not clearly visible are hard to assess for the classifier and therefore more prone to misclassification.

These observations are valid for many of the attributes but there are attributes, like backpack, for which different results are received. As an example, easy and hard samples for the attribute *Backpack* are shown in Figure 4.4. All easy samples show persons without a backpack whereas each of the persons from the particularly hard samples wears a backpack. So, in this case it seems that the decision between hard and easy images is only taken based on the presence of

(a) Easy samples

(b) Hard samples

**Figure 4.3**: Hard and easy samples for the attribute *Gender* of the RAP-2.0 dataset based on the estimated hardness scores. Samples that are considered easy or hard appear to be reasonable for this attribute.

the attribute and by that equals the classifier instead of providing independent hardness predictions. This indicates that, albeit the hardness predictor loss is weighted by the positive ratio of attributes, the imbalance of attributes in the training data still plays a big role and influences the recognition accuracy negatively. Since only about 1 % of the training images show persons with backpacks, the network can achieve good results by only predicting no backpack. Thus, the loss gets minimal for such images and high for images with backpacks. As a result, the hardness predictor learns to discriminate between the values of the attribute and not to predict the hardness of the attribute recognition task.

## 4.4 Realistic prediction

Based on the finding that the hardness predictor can give meaningful estimates of the degree of difficulty of samples, the realistic predictor can be formed by combining the classifier with a hardness-based rejection. Table 4.2 presents

(a) Easy samples

(b) Hard samples

**Figure 4.4**: Hard and easy samples for the attribute *Backpack* of the RAP-2.0 dataset based on the estimated hardness scores. In contrast to *Gender*, persons with backpacks are considered hard-to-classify due to the high attribute imbalance.

the results for different rejection strategies and compares them to confidence score-based rejection. Improvements in instance-based metrics can be observed, independent of the applied rejection method. The mA-score decreases except for the mA rejection. This is due to the side effects of unbalanced attributes, which are always predicted as false and thus reach only a minimum mA score of $0.5$. When comparing rejection methods, threshold strategy achieves the best F1 scores whereas, as mentioned above, mA rejection leads to highest mA results. Although hardness prediction-based rejection of attributes increases the performance, rejection on the basis of class probabilities achieves similar or even better performance, especially on RAP-2.0 dataset. This finding indicates that the major issue with the external hardness prediction network is still the unbalanced distribution of attribute values and that DeepMAR weighted loss function is not completely capable of compensating it.

In conclusion, it can be stated that the realistic predictor approach using an external hardness predictor generally works. But the assumption that such an additional CNN is superior to the use of confidence scores cannot be fully validated for the pedestrian attribute task. Both networks learn complementary

**Table 4.2**: Realistic predictor results on RAP-2.0 dataset. Rejection strategies mainly improve instance metrics. Hardness scores provided by an explicit hardness predictor do not surpass the baseline given by using confidence scores of the classifier.

| | RAP2.0 | | | PA-100K | | |
|---|---|---|---|---|---|---|
| Rejection Strategies | mA | F1 | Rejected | mA | F1 | Rejected |
| None | 72.98 | 77.12 | 0.00 | 75.23 | 83.33 | 0.00 |
| *Hardness scores:* | | | | | | |
| Threshold | 69.18 | 83.59 | 12.54 | 74.34 | 90.53 | 15.04 |
| Quantile | 66.07 | 81.80 | 24.58 | 74.20 | 88.08 | 22.54 |
| mA | 74.02 | 78.93 | **7.75** | **78.09** | 90.32 | 15.18 |
| F1 | 66.14 | 79.52 | 16.68 | 74.78 | 87.67 | **13.44** |
| *Confidence scores:* | | | | | | |
| Threshold | 71.77 | **85.98** | 14.51 | 75.87 | **91.20** | 17.32 |
| mA | **74.79** | 82.88 | 12.13 | 77.88 | 91.00 | 17.44 |

tasks and so the rejection rate is much lower when the hardness predictor network is used. However, results of the confidence score are not exceeded.

# 5 Conclusion and future work

This work aimed to apply the concept of realistic predictors to the field of pedestrian attribute recognition. The core idea was to address some of the biggest challenges in pedestrian attribute recognition while simultaneously achieving more reliable attribute estimates. To achieve this, the approach introduced in [17] was modified and optimized for the task of attribute recognition. This included, for instance, adapting the loss functions and alterations regarding to the network architecture. In addition, different strategies to determine meaningful thresholds for exclusion of unreliable predictions were proposed and extensively studied.

All in all the findings of this work showed that the concept of realistic predictors can be transferred to the field of pedestrian attribute recognition and accuracy improvements can be achieved. However, comprehensive experiments indicate

that the predictions of hardness do not reflect the difficulty of the task equally well for all attributes. Especially attributes with strongly unbalanced value distributions in the training dataset cause problems and worsen the results. As a result, better performance was achieved if confidence scores are used instead of hardness predictions. In one point, however, the hardness predictions were strongly superior to the confidence values, namely in the number of rejected samples. From this it can be concluded that training a separate hardness predictor has its advantages.

In future research the training of the hardness predictor and the loss function can be improved in order to eliminate the imbalance problem of some attributes. The aim is to close the performance gap with the confidence-based rejection while maintaining the advantage in terms of number of rejected samples. Moreover, the hardness predictor approach allows to weight attributes during attribute-based person retrieval. By considering attributes according to their difficulty in predicting them during distance computation, incorrect retrieval results in early ranking positions can be avoided.

# References

[1] Abrar H Abdulnabi et al. “Multi-task CNN model for attribute prediction”. In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 1949–1959.

[2] Esube Bekele and Wallace Lawson. “The Deeper, the Better: Analysis of Person Attributes Recognition”. In: *arXiv preprint arXiv:1901.03756* (2019).

[3] Ali Diba et al. “Deepcamp: Deep convolutional action & attribute mid-level patterns”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3557–3565.

[4] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. “Actions and attributes from wholes and parts”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2470–2478.

[5] Hao Guo, Xiaochuan Fan, and Song Wang. “Human attribute recognition by refining attention heat map”. In: *Pattern Recognition Letters* 94 (2017), pp. 38–45.

[6] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[7] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. “Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios”. In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE. 2015, pp. 111–115.

[8] Dangwei Li et al. “A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios”. In: *IEEE transactions on image processing* 28.4 (2018), pp. 1575–1590.

[9] Dangwei Li et al. “Pose guided deep model for pedestrian attribute recognition in surveillance scenarios”. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2018, pp. 1–6.

[10] Yining Li et al. “Human attribute recognition by deep hierarchical contexts”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 684–700.

[11] Pengze Liu et al. “Localization guided learning for pedestrian attribute recognition”. In: *arXiv preprint arXiv:1808.09102* (2018).

[12] Xihui Liu et al. “Hydraplus-net: Attentive deep features for pedestrian analysis”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 350–359.

[13] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. “Deep imbalanced attribute classification using visual attention aggregation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 680–697.

[14] M Saquib Sarfraz et al. “Deep view-sensitive pedestrian attribute inference in an end-to-end model”. In: *arXiv preprint arXiv:1707.06089* (2017).

[15] Arne Schumann, Andreas Specker, and Jürgen Beyerer. “Attribute-based Person Retrieval and Search in Video Sequences”. In: *Advanced Video and Signal Based Surveillance (AVSS), 2018 15th IEEE International Conference on*. 2018.

[16] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. “Person attribute recognition with a jointly-trained holistic cnn model”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 87–95.

[17] Pei Wang and Nuno Vasconcelos. “Towards realistic predictors”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 36–51.

[18] Luwei Yang et al. “Attribute recognition from adaptive parts”. In: *arXiv preprint arXiv:1607.01437* (2016).

# Towards a Formal Model for Quantifying Trust in Distributed Usage Control Systems

*Paul Georg Wagner*

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

Technical Report IES-2019-06

## Abstract

Distributed usage control is a form of usage control that spans over multiple domains and computer systems. As a result, usage control components responsible for evaluating policies, gathering information, executing actions and enforcing decisions are operated in the vicinity of different stakeholders with conflicting interests. In order to prevent malicious stakeholders from manipulating these components, remote attestation can be used to verify the integrity of their code base. However, in a distributed case it is not always apparent what sequence of attestations is necessary and which verifier should conduct them. Furthermore, it is unclear what impact a failed attestation has on the trustworthiness of the whole usage control system. To solve these questions, it is necessary to identify which agents need to trust each other in order to securely execute a certain usage control function. Then the sequence of remote attestations that occur across the distributed usage control system can be examined accordingly. In this work we develop a formal model that represents the trust relationships of distributed usage control systems with multiple collaborating actors. Based on the conducted attestations we define simple binary and non-binary trust metrics that quantify the trust level a data owner can expect at a certain point in time.

Finally we show how the model can be used to determine the level of trust reached in a real-world scenario.

# 1 Introduction

In recent years, usage control (UC) has been more and more propagated as a novel technology for governing access to valuable information. Unlike classical access control, usage control models focus on managing the future usage of data [7]. With usage control technology it is possible to restrict access to protected assets even after they have been disclosed. Often usage control is used in distributed environments, where sensitive data are shared between shareholders. One such example is the Fraunhofer research project *International Data Space* [6]. The International Data Space allows data providers to distribute valuable data alongside usage restrictions to potentially malicious data consumers. The data consumer's systems then process the received information according to the published rules. Naturally, the data provider wants to ensure that the data consumer can be trusted to obey the issued usage restrictions on his data. For this the International Data Space uses distributed UC modules that independently evaluate the usage control policies and enforce the resulting decisions. Since each participant of the data space may act maliciously and try to extract foreign data past the protection mechanisms, it is necessary to verify the integrity of the UC components prior to the data exchange.

*Trusted computing* is the state of the art approach that allows for remote verification of software components. Currently the most widespread trusted computing technologies are Trusted Platform Modules (TPMs) [9] and Intel's Software Guard Extensions (SGX) [3]. Both of these technologies support establishing trust in remote software stacks by verifying code bases through special hardware and cryptographic methods. This software verification process is called *remote attestation*. Besides verifying the integrity of a software stack, remote attestation also establishes secure channels between prover and verifier. The International Data Space uses TPMs and a customized remote attestation protocol to establish trust in data consumers. However, when developing distributed usage control systems that establish trust by remote attestation,

several open questions remain. For example, it is not always clear which components have to be attested, and by which verifiers. Comprehensive usage control systems are complicated and security relevant UC functions may span over multiple distributed UC components. This is especially true if the usage control system also includes components that track and store the provenance of supervised data. In these cases it has to be ensured that all involved UC components are properly attested and can securely communicate with each other. Another interesting question is what impact a failed attestation has on the security of the overall system. These questions all emerge from the yet unsolved problem of quantifying the trust propagation in dynamically operating and distributed usage control systems.

In this work we develop a formal model that can represent the trust relationships that occur in distributed usage control systems with multiple collaborating actors. This model is independent from the design of the UC-system, its implementation, and the used trusted computing technology. Furthermore we define simple binary and non-binary trust metrics that can be used to determine the trust level of certain UC functionalities at a specific point in time. Calculating a dynamic trust level for a UC system is very beneficial for conducting a comprehensive security analysis of the infrastructure. Finally we show how the model can be used to determine the level of trust in a real-world example scenario based on the International Data Space using TPM-based attestation.

# 2 Related work

Managing and distributing trust has been a major topic of research interest for a long time. By far the most widespread technique of managing trust in distributed systems is via a *public key infrastructure* (PKI). With a PKI, a few trusted certification authorities (CAs) issue signed public keys for the agents in their domain. As a result, the trust in a certain communication channel is reduced to the trustworthiness of the CA. Even though PKIs are a fundamentally important concept in IT security, as a centralized way of managing trust they are not applicable to our scenario. In terms of decentralized approaches to trust management, the most important concept is the *Web of Trust* [1], which has

been popularized by the well-known PGP software. Its main principle is to distribute trust transitively by endorsement of already trustworthy collaborators (i.e. “my friend’s friend is my friend”). Also, it is possible to generate new trust by offline comparison of public key fingerprints. This decentralized version of trust distribution already comes close to what we need for our scenario. A usage control component could determine the level of trust in a remote system based on the trust that their peers already have in it. New trust would then be generated by automated remote attestation instead of manually comparing fingerprints. However, the Web of Trust does not offer any kind of trust metric, and does not take possible internal attackers into account. Also it does not give any notion of time.

An approach that factors in these aspects are dynamic reputation systems [5, 4, 10]. Their idea is to describe trust mathematically and develop a metric for the reputation of an agent based on their previous behavior. Simply put, if an agent behaves cooperatively, its level of trust increases. If the agent defects, the trust level is impacted. However, since it is not at all well-defined what constitutes as “cooperative behavior” in our scenario, reputation systems also do not suffice for quantifying trust in distributed UC systems. Furthermore, they neither define what actions are suitable to increase or decrease trust, nor do they deal with attestation mechanisms. Since our goal is to develop a formal model of distributed UC systems that works independently of the system design or the used attestation technology, reputation systems do not meet our requirements.

# 3 Formal model

Our goal is to develop a metric that quantifies the level of trust in distributed UC systems. For this, a formal model is required that describes the trusted communication between usage control components. Since trust relationships can be intuitively modeled as graphs, we utilize a graph-based approach. Furthermore, the formal model needs to represent attestations conducted by the UC components as well as the architecture of the deployed UC system. In this section, we develop a suitable model in three steps.

1. Define functions of the UC components that have to be trusted using a graph-based model (global).
2. Define the existing agents and cross-system activities of interest by instantiating that graph (scenario specific).
3. Define the system architecture by binding the agents to attestable systems (implementation specific).

As a first step, the basic semantic of the usage control system is specified via a *trust dependency graph*. The trust dependency graph contains the existing types of usage control components. It describes how they need to trust each other for any interaction that may occur between them. In the second step we concretize this graph by considering the actual components that are operated in the distributed usage control system. For this we represent each concrete UC component as an instance of a node from the trust dependency graph. We call a concrete UC component *agent*, because it needs to securely interact with other components in the system. The resulting graph is called *agent graph*. Unlike the trust dependency graph, each agent graph is specific to a certain scenario that the UC system is deployed for. It also yields information about the actors that operate the usage control components in that scenario. The agent graph can be partitioned into multiple *UC activities*, which represent a function of the distributed UC system spanning over multiple UC components. We will later show how the trust level of a UC activity can be measured using an instance of the model. Finally, an *architecture graph* defines how the agents map to physical computer systems that can be attested. The architecture graph is not only specific for a certain UC scenario, but also depends on the used trusted computing technology and the deployment of UC components. Figure 3.1 shows an overview of the steps required to transfer the design and implementation of a UC system into the formal model. In the following sections we present this formal model in detail. Afterwards we develop trust metrics that can be evaluated on an instance of the model.

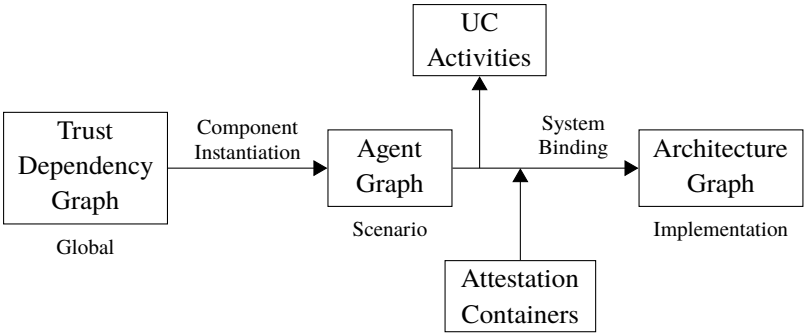



**Figure 3.1**: Overview of the formal model.

## 3.1 Defining UC systems

The first step of the formal model includes defining a distributed UC system and its components. This is done in definition 3.1.1.

**Definition 3.1.1** (DUC system)**.** Let $M$ be a finite set of DUC modules and $F$ a set of DUC functions. We call the tuple $S := (M, F)$ *DUC system*.

Besides the UC components and their functions, we also need to define how the UC components may interact with each other. This is done by the trust dependency graph, as described in definition 3.1.2.

**Definition 3.1.2** (Trust Dependency Graph)**.** Let $S = (M, F)$ be a DUC system. Let $E^F \subseteq M \times M$ be a set of directed edges over $M$ and $l^F : E^F \to F$ a mapping that labels each edge with a system function. We call the triple $FG := \left(M, E^F, l^F\right)$ *trust dependency graph* of $S$.

The trust dependency graph of a DUC system describes the inter-component functions that may be called across the distributed system. A trust dependency graph can be constructed solely with knowledge of the UC component's interfaces. It is not necessary to know the use case or the usage control policies that should be deployed. Hence the trust dependency graph is independent of the system's concrete realization and implementation.

An example for a trust dependency graph is presented in figure 3.2. It shows the trust dependency graph for the XACML-based distributed usage control architecture that is deployed in the International Data Space. XACML [2] is a reference architecture that defines usage control components responsible for enforcement (PEP), policy evaluation (PDP), information gathering (PIP), and administration (PAP). Besides these XACML-based components, the usage control architecture of the International Data Space uses some additional components responsible for retrieving policies (PRP), managing communication (PMP) and executing obligations (PXP). The displayed DUC system is modeled as $M = \{PEP, PDP, PIP, ...\}$ and $F = \{notify, evaluate, execute, ...\}$. The trust dependency graph shows the possible interactions and the resulting trust dependencies between components as labeled edges. Note that the direction of

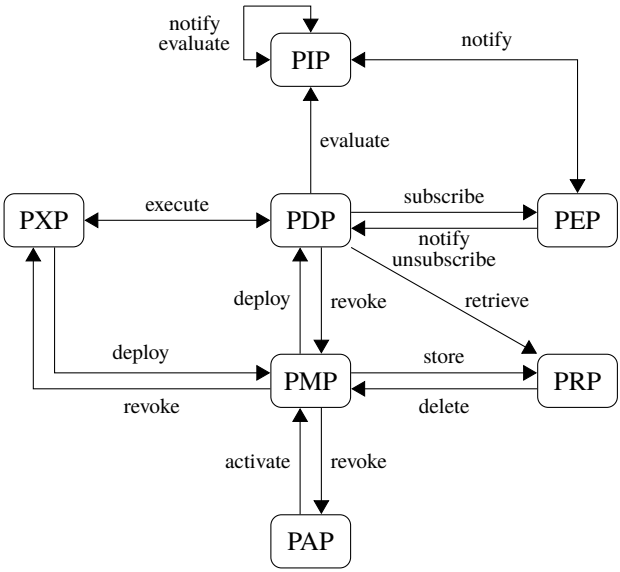



**Figure 3.2**: Example of a trust dependency graph.

the edge defines the direction of the trust dependency, which does not always correspond with the direction of the interaction. For example, a PAP may revoke

a policy by calling the *revoke* function of the responsible PMP. However, the edge is directed the opposite way, because in this case the PMP has to trust the PAP that the revocation request is legit.

## 3.2 Defining agents and activities

In order to represent a specific scenario, we can instantiate the trust dependency graph and introduce agents that interact with each other. This is done in definition 3.2.1.

**Definition 3.2.1** (Agent Graph)**.** Let $FG = \left(M, E^F, l^F\right)$ be a trust dependency graph. Let $A$ be a set of agents, $E \subseteq A \times A$ a set of directed edges over $A$ and $l : A \to F$ a mapping. Let also be $type : A \to M$ a mapping that assigns a module type to each agent. We call the tuple $G := (A, E, l, type)$ *agent graph*, if it holds that

$$\forall (a, b) \in E : (type(a), type(b)) \in E^F$$
$$\forall (a, b) \in E : l\,(a, b) = l^F\,(type(a), type(b))$$

According to definition 3.2.1, every agent is an instance of a UC component. The agent interaction corresponds to the DUC functions that have been described by the trust dependency graph. The two conditions in 3.2.1 ensure that the agent graph only contains edges that correspond to the trust dependency graph (i.e. agents can only call existing functions). Note that the agent graph may contain multiple agents of one particular type (e.g. if multiple PIPs or PEPs exist), while the trust dependency graph contains each component exactly once.

The agent graph shown in figure 3.3 is based on the example trust dependency graph in figure 3.2. The example agent graph shows a scenario with two actors A and B, who operate distributed usage control components. In this scenario, the PXP instance of actor B is responsible for deploying policies at the PDP instance of actor A. This allows B to enforce usage control policies on his data, even if they are shared with A. Note that the agent graph contains multiple instances of a single UC component. For example, in this case both actors A and B operate PDPs, PEPs and PXPs. While the trust dependency graph is of a global nature and represents an abstract DUC architecture, agent graphs

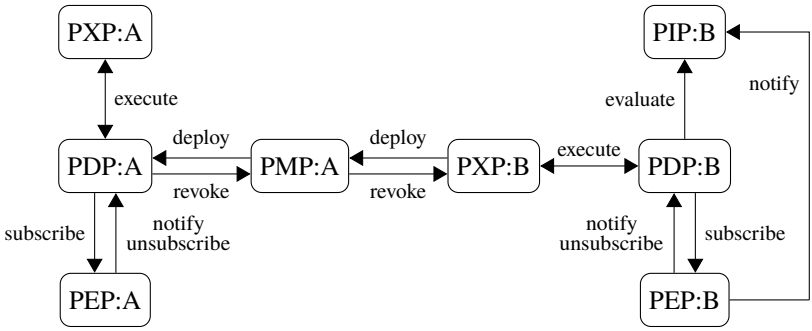



**Figure 3.3**: Example of an agent graph.

derived from it depend on specific use cases. Also note that all agent graph edges correspond to edges from the trust dependency graph, but not all trust relationships may be included in the agent graph, depending on their relevance for the scenario.

Besides defining the agents of the UC system, we also have to specify what kind of agent interaction should be evaluated for trustworthiness. Definition 3.2.2 partitions the agent graph into multiple acyclic subgraphs called *UC activities*. A UC activity represents an action that requires multiple agents to work together, such as the deployment of policies or the enforcement of access decisions. Since the involved agents have to trust each other in order to reliably execute these actions, the trust level of a UC system will be based on the relevant UC activities.

**Definition 3.2.2** (UC Activity)**.** Let $G = (A, E, l, type)$ be an agent graph. Let $H := (\bar{A}, \bar{E}, \bar{l})$ be a connected subgraph of $G$ with $\bar{A} \subseteq A$, $\bar{E} \subseteq E \cap \left(\bar{A} \times \bar{A}\right)$ and $\bar{l} := l|_{\bar{E}}$. We call the subgraph $H$ *UC activity* of $G$, if

$$H \text{ is acyclic}$$
$$\exists! x \in \bar{A} : indeg(x) = 0$$
$$\exists y \in \bar{A} : outdeg(y) = 0$$

The unique vertex $x$ is called *root* of $H$. A vertex $y$ is called *leaf* of $H$. The set of all leaves is denoted by $Y$.

Figure 3.4 shows an example for a UC activity based on the agent graph in figure 3.3. The depicted UC activity represents the necessary interaction for locally enforcing a policy. First, the enforcement point (PEP) notifies the decision point (PDP) of an access request. The PDP then evaluates the policies, requests necessary information at the PIP and executes obligations at the PXP. In this activity the PEP acts as root, while the PXP and the PIP are leaves. In order to trust the UC activity of local policy enforcement, all of these interactions need to be secure. Complex distributed usage control systems, such as the International Data Space, have many more relevant UC activities that can be identified, including remote policy enforcement, policy deployment, and policy revocation. However, for the remainder of this paper we will stick to the example of local policy enforcement.

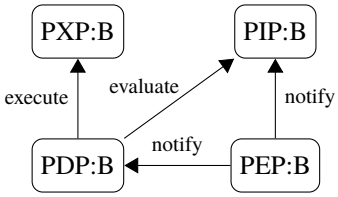



**Figure 3.4**: Example of a UC activity: Local policy enforcement.

## 3.3 Defining attestations and architectures

Finally the formal model needs to contain information about the remote attestations that can be executed by the agents. In order to accommodate this, definition 3.3.1 introduces the notion of *attestation containers*. An attestation container is a set of agents that can be jointly attested. Which agents form an attestation container depends on the used attestation technology and the system architecture. For example, if the UC system uses TPMs to execute the remote attestations, all UC components running on a TPM-protected computer system are included in an attestation container. More advanced trusted computing technologies, such as Intel's SGX, allow the attestation of software enclaves rather than whole

computer systems. In that case, all UC components included inside such an enclave form an attestation container. The tuple of agent graph and attestation containers is called *architecture graph*.

**Definition 3.3.1** (Attestation Container)**.** Let $G = (A, E, l, type)$ be an agent graph. We call a non-empty set $C \subseteq A$ *attestation container*, if all $c \in C$ can be jointly attested. The set of all attestation containers is denoted by $\mathcal{C} \subseteq \mathcal{P}(A) \setminus \emptyset$. The tuple $(G, \mathcal{C})$ is called *architecture graph*.

Based on the description of the attestation container, we have to represent the concrete attestations that agents actually conduct during runtime. This is done in definition 3.3.2 via an *attestation schedule*.

**Definition 3.3.2** (Attestation Schedule)**.** Let $(G, \mathcal{C})$ be an architecture graph. For any agent $a \in A$ we call the mapping $att_a : \mathbb{N}^+ \times \mathcal{C} \to \{-1, 0, 1\}$ *attestation schedule*. The family of all attestation schedules is denoted by $\mathcal{A} = (att_a)_{a \in A}$.

The attestation schedule of an agent indicates which attestations the agent conducts at what points in time, and if they are successful. More concretely, if $att_a(t, C) = 1$, then at time $t$ the agent $a$ conducts a successful remote attestation of container $C$. This means that $a$ successfully verifies the integrity of all agents that are included in $C$. If instead $att_a(t, C) = -1$, the attestation fails and the agent is unable to verify the integrity of $C$. If $att_a(t, C) = 0$, the agent $a$ does not conduct a remote attestation of container $C$ at time $t$.

# 4 Quantifying trust

The formal model allows us to mathematically represent a distributed UC system. Based on an architecture graph and the associated attestation schedules we can now define trust metrics for the relevant UC activities.

## 4.1 Binary trust metrics

Given a UC activity $H$, we denote the level of trust in the activity at time $t$ by $TrustLevel^H(t) \in \{0, 1\}$. A trust level of 1 means that the activity is trusted,

while a trust level of 0 indicates that the attestations are not sufficient to ensure the integrity of all involved components. In order to define this trust level, we are examining the paths of $H$ and calculate trust gains for each transition within the path.

### 4.1.1 Trust gain by attestations

Whenever an agent $a$ conducts a successful attestation of container $C$, possible transitions between $a$ and another agent $c \in C$ are trusted and positively influence the trust level of $H$. However, this positive influence only lasts as long as no other agent unsuccessfully conducts an attestation of $C$, thereby determining that its integrity cannot be trusted anymore. This idea is expressed in definition 4.1.1. Like the overall trust level, the trust gain is binary. A trust gain of 1 for the transition $(a, b)$ means that $b$ has been attested by $a$, and no other agent failed in verifying the integrity of $b$ since. A trust gain of 0 indicates that $a$ has not yet attested a container that includes $b$, or that such an attestation is outdated.

**Definition 4.1.1** (Trust Gain by Attestation)**.** Let $(G, \mathcal{C})$ be an architecture graph and $\mathcal{A} = (att_a)_{a \in A}$ the family of associated attestation schedules. Let $H$ be a UC activity of $G$ and $(v_1, ..., v_n) \in H$ a path of the activity. The trust gain by attestation for the transition $(v_{i-1}, v_i)$ at time $t$ is defined as

$$Gain^{att}(i,t) := \begin{cases} & \exists C \in \mathcal{C}, t_1 \leq t : \\ 1, & \quad v_i \in C \wedge att_{v_{i-1}}(t_1, C) = 1 \wedge \\ & \quad \forall a \in A : \nexists t_2 \in [t_1, t] : att_a(t_2, C) = -1 \\ 0, & \text{else} \end{cases}$$

### 4.1.2 Trust gain by locality

While it is clear that attesting a remote component increases trust, we also have to manage the trust gains of local components. If two dependent UC components are included in the same attestation container, they can communicate securely without conducting a remote attestation. However, even though a remote attestation is not required for establishing a secure channel, the integrity of both components still needs to be verified. Hence we have to demand that a previous

component attests both of the local components. This concept of trust gain by locality is specified in definition 4.1.2.

**Definition 4.1.2** (Trust Gain by Locality)**.** Let $(G, \mathcal{C})$ be an architecture graph and $\mathcal{A} = (att_a)_{a \in A}$ the family of associated attestation schedules. Let $H$ be a UC activity of $G$ and $(v_1, ..., v_n) \in H$ a path of the activity. The trust gain by locality for the transition $(v_{i-1}, v_i)$ at time $t$ is defined as

$$Gain^{loc}(i,t) := \begin{cases} \exists C \in \mathcal{C}, t_1 \leq t : \\ \quad 1, \quad \{v_{i-1}, v_i\} \subseteq C \wedge \\ \quad \exists j < i : att_{v_j}(t_1, C) = 1 \wedge \\ \quad \forall a \in A : \nexists t_2 \in [t_1, t] : att_a(t_2, C) = -1 \\ 0, \text{ else} \end{cases}$$

### 4.1.3 Putting it together

Given the two concepts of generating trust in a distributed UC system, we can define the trust level for a UC activity. We can base the definition on the trust gain by attestation, the trust gain by locality, or both. Definition 4.1.3 specifies the trust level of a path by multiplying the trust gains of the respective transitions. The trust level of the whole UC activity is the minimal trust over all paths.

**Definition 4.1.3** (Trust Level)**.** Let $(G, \mathcal{C})$ be an architecture graph and further let $H = (\bar{A}, \bar{E}, \bar{l})$ a UC activity of $G$ with root $x \in \bar{A}$ and leaves $Y \subseteq \bar{A}$. The trust level of a path $(v_1, ..., v_n) \in H$ is defined as

$$TrustLevel^{(v_1,...,v_n)}(t) := \prod_{i=2}^{n} (Gain(i,t))$$

Depending on the scenario, the trust gain is defined by attestation or attestation and locality.

$$Gain(i,t) := Gain^{att}(i,t)$$
$$Gain(i,t) := \max(Gain^{att}(i,t), Gain^{loc}(i,t))$$

The trust level of the UC activity $H$ is defined as

$$TrustLevel^{H}(t) := \min_{\substack{(v_1,...,v_n) \in H \\ v_1 = x, v_n \in Y}} \Big( TrustLevel^{(v_1,...,v_n)}(t) \Big)$$

Note that the trust level definition is based on the transitions between agents in the UC activity, instead of the agents themselves. Unlike many existing reputation systems (c.f. section 2), we do not define the trust level of a certain agent at all. Instead we define the trust gain of a transition within a UC activity, and then generalize that definition over paths to the whole activity. The reason for this is that remote attestation is not only responsible for verifying the integrity of agents, but also establishes a secure channel for communication. Hence it is not sufficient to focus just on the level of trust in the agent, we need to examine the connections between them.

## 4.2 Non-binary trust metrics

A binary trust metric can only distinguish trusted from untrusted systems. In order to quantify trust more precisely, we can define non-binary trust metrics. In that case, given a UC activity $H$, we denote the level of trust in the activity at time $t$ by $TrustLevel^H(t) \in [0, 1]$.

A simple non-binary trust metric can be obtained by including the temporal decay of trust in the model. For this we introduce a dampening factor $\eta : \mathbb{N}_0^+ \to [0, 1]$ and modify the definitions of trust gains.

**Definition 4.2.1** (Trust Gains with Temporal Decay)**.**

$$Gain^{att}(i,t) := \begin{cases} \eta(t-t_1), & \exists C \in \mathcal{C}, t_1 \leq t : \\ & \quad v_i \in C \wedge att_{v_{i-1}}(t_1, C) = 1 \wedge \\ & \quad \forall a \in A : \nexists t_2 \in [t_1, t] : att_a(t_2, C) = -1 \\ 0, & \text{else} \end{cases}$$

$$Gain^{loc}(i,t) := \begin{cases} \eta(t-t_1), & \exists C \in \mathcal{C}, t_1 \leq t : \\ & \quad \{v_{i-1}, v_i\} \subseteq C \wedge \\ & \quad \exists j < i : att_{v_j}(t_1, C) = 1 \wedge \\ & \quad \forall a \in A : \nexists t_2 \in [t_1, t] : att_a(t_2, C) = -1 \\ 0, & \text{else} \end{cases}$$

The definition of the dampening factor $\eta$ depends on the scenario. In general, the choice of $\eta$ reflects how fast the generated trust deteriorates after a successful

attestation. For most cases a polynomial or exponential decay should be an adequate choice.

$$\eta(t) := (t+1)^{-p}$$
$$\eta(t) := \exp(-\lambda t)$$

## 4.3 Example calculation

After defining binary and non-binary trust metrics, we give an example calculation based on the previously used International Data Space scenario. For this, we take the UC activity representing local policy enforcement from figure 3.4 and define suitable attestation containers. As shown in figure 4.1, the set of attestation containers results to $\mathcal{C} = \{\{pip\}, \{pdp, pxp\}\}$. Since the International Data Space uses TPMs to provide proof of integrity during remote attestation, in this case the attestation containers represent physical computer systems.

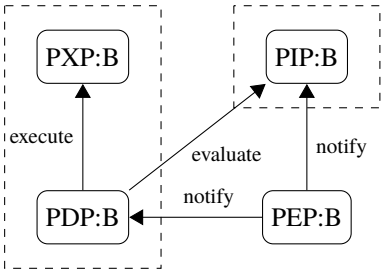



**Figure 4.1**: UC activity with attestation containers: Local policy enforcement.

In order to determine the trust level of this scenario, we have to specify the attestation schedule. We assume that the PXP and the PIP do not conduct any attestations in this example.

$$\forall t \in \mathbb{N}^+, C \in \mathcal{C} : att_{pxp}(t, C) = 0$$
$$\forall t \in \mathbb{N}^+, C \in \mathcal{C} : att_{pip}(t, C) = 0$$

The root PEP attests the PDP and PXP at $t = 1$, while both PEP and PDP attest the PIP at $t = 2$. Since in this example all conducted attestations are successful, the attestation schedules never evaluate to $-1$.

$$att_{pep}(t, C) = \begin{cases} 1, \ t = 1 \land C = \{pdp, pxp\} \\ 1, \ t = 2 \land C = \{pip\} \\ 0, \ \text{else} \end{cases}$$

$$att_{pdp}(t, C) = \begin{cases} 1, \ t = 2 \land C = \{pip\} \\ 0, \ \text{else} \end{cases}$$

Furthermore we consider both attestation and locality trust gains for this example calculation. Table 4.1 shows the development of the trust level for the three paths.

| $t$ | $TL^{(pep,pdp,pxp)}$ | $TL^{(pep,pdp,pip)}$ | $TL^{(pep,pip)}$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | $\eta^{att}(0) * \eta^{loc}(0)$ | 0 | 0 |
| 2 | $\eta^{att}(1) * \eta^{loc}(1)$ | $\eta^{att}(1) * \eta^{att}(0)$ | $\eta^{att}(0)$ |

**Table 4.1**: Development of trust levels over time.

At $t = 0$, no attestations have been conducted yet, so the trust level for all paths is 0. At $t = 1$, the PEP conducts a remote attestation of the attestation container $\{pdp, pxp\}$. This results in an attestation trust gain of $\eta^{att}(0)$ for the transition $pep \rightarrow pdp$ and a locality trust gain of $\eta^{loc}(0)$ for the transition $pdp \rightarrow pxp$. At $t = 2$, both the PEP and the PDP conduct a remote attestation of the attestation container $\{pip\}$. Then the transition $pep \rightarrow pip$ is directly attested with an attestation trust gain of $\eta^{att}(0)$. However, the transition $pep \rightarrow pdp$ now has an attestation trust gain of $\eta^{att}(1)$, since the relevant attestation is one time step in the past. For the same reason the transition $pdp \rightarrow pxp$ now has a locality trust gain of $\eta^{loc}(1)$.

If we assume the dampening factors of the trust gains to be

$$\eta^{att}(t) := \exp(-\frac{1}{10}t)$$

and $\eta^{loc}(t) := \exp(-\frac{1}{15}t)$, the trust level of the entire activity $H$ at time $t = 2$ results to

$$\begin{aligned} \mathit{TrustLevel}^H(2) &= \min\left(\eta^{att}(1) * \eta^{loc}(1), \eta^{att}(1) * \eta^{att}(0), \eta^{att}(0)\right) \\ &= \min\left(\eta^{att}(1) * \eta^{loc}(1), \eta^{att}(1) * 1, 1\right) \\ &= \eta^{att}(1) * \eta^{loc}(1) \\ &= 0.846 \end{aligned}$$

# 5 Conclusion

In this work we developed a formal model for quantifying trust in distributed usage control systems. After defining the relevant trust dependencies and interacting agents, we developed binary and non-binary trust metrics that quantify the level of trust reached in a certain scenario. While successful attestations positively influence the trust, failed attestations and time progression reduce the reached overall trust level. Finally we showed an example calculation based on the real distributed usage control system that is deployed in the International Data Space.

Possible future work includes investigating how Dempster-Shafer theory [8] could be applied to the formal model. With Dempster-Shafer it is possible to model unawareness and uncertainty of knowledge. It is also helpful in combining degrees of belief from different sources, which makes it promising for representing trust in distributed systems. There already are reputation systems based on Dempster-Shafer theory [12].

Another important approach is to evaluate to what extent the assumptions made by the formal model hold in practice. The presented trust metric is only meaningful if the used remote attestation protocol guarantees integrity verification and secure communication across the distributed system. However, especially for the widespread TPMs this assumption does not hold in all scenarios [11]. A more subtle problem that occurs in practice is the availability of UC components.

Even if the used remote attestation protocol is secure, one can never prevent a malicious operator to deliberately sever communications between local and remote usage control components. In this case it is important that the roots of all affected UC activities are notified about the loss of communication, otherwise the security of the usage control system may be compromised. Even though the formal model cannot directly monitor this, being able to identify relevant UC activities and their trust dependencies is a substantial help in auditing distributed usage control systems for these weaknesses.

# References

[1] Alfarez Abdul-Rahman. “The pgp trust model”. In: *EDI-Forum: the Journal of Electronic Commerce*. Vol. 10. 3. 1997, pp. 27–31.

[2] Anne Anderson et al. “extensible access control markup language (xacml) version 1.0”. In: *OASIS* (2003).

[3] Victor Costan and Srinivas Devadas. “Intel SGX Explained”. In: *IACR Cryptology Archive* (2016), p. 86.

[4] Audun Josang and Roslan Ismail. “The beta reputation system”. In: *Proceedings of the 15th bled electronic commerce conference*. Vol. 5. 2002, pp. 2502–2511.

[5] Stephen Paul Marsh. “Formalising trust as a computational concept”. In: (1994).

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

[7] 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.

[8] Glenn Shafer. *A mathematical theory of evidence*. Vol. 42. Princeton university press, 1976.

[9] TCG. “Architecture overview”. In: *Specification Revision* 1 (2007).

[10] H Vagts, T Cosar, and J Beyerer. “Establishing trust in decentralized smart sensor networks”. In: *Mobile Multimedia/Image Processing, Security, and Applications 2011*. Vol. 8063. International Society for Optics and Photonics. 2011, p. 806306.

[11] 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.

[12] Bin Yu and Munindar P Singh. “An evidential model of distributed reputation management”. In: *Proceedings of the first international joint conference on Autonomous Agents and Multiagent Systems: Part 1*. ACM. 2002, pp. 294–301.

# Learning with Latent Representations of 3D Data: from Classical Methods to 3D Deep Learning

*Chengzhi Wu*

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

Technical Report IES-2019-11

## Abstract

3D data contain rich information about the full geometry of objects or scenes. Learning tasks on them have always been considered as hard ones in the computer vision community due to their extreme high dimensionality. Hence, latent representations of 3D geometries are often used to lower the data dimensionality for better parameterization and easier computation. In this report, we make a brief review on those latent representations obtained via different methods including classical ones and the emerging neural learning-based ones. Furthermore, the nowadays widely used deep learning methods have also been more closely investigated regarding their applications on various 3D data formats. The possibility of combing those two kinds of methods has also been addressed.

## 1 Introduction

3D data analysis has always been an interesting yet challenging research topic for computer vision researchers. Learning latent information from them is vital

to lots of advanced technology applications including robotics, autonomous driving, virtual reality and augmented reality. Lots of classical methods have been proposed to extract latent representations from 3D data. Those latent representations can be images, graphs, histograms, or even vectors [4]. Classical methods usually focus more on generating latent representations of 3D shapes. Those generated latent representations are sometimes also referred as shape descriptors.

In recent years, neural networks have been proved to be one of the most powerful learning algorithms for computer vision tasks, especially on 2D Euclidean data. Implicitly learned feature maps or bottleneck feature vectors have been used for classification, detection, or segmentation tasks. Later on, similar methods have been proposed on 3D Euclidean data with minor adaptions. However, those learning algorithms cannot be straightforwardly extended to Non-Euclidean data due to their non-grid data structure. Different special neural network architectures for 3D Non-Euclidean data therefore have been more meticulously designed and proposed, while input, output, latent representations, or even network operations have been more artfully defined.

This report is structured as follows. In Section 2, we briefly review the most common 3D data formats. Latent representations learned by classical methods or neural learning-based methods are reviewed in Section 3. Section 4 gives a more detailed review on the application of deep learning a) for ML tasks on 3D data and b) for the generation of latent representations that can be used by different methods later on. Conclusion and future outlook are presented in Section 5.

# 2 Overview of 3D data format

3D data have lots of different formats depending on its source. They are usually categorized into 2 subsets, Euclidean data, which mainly include multi-view images, RGB-D images, volumetric voxels or octrees; and Non-Euclidean data, which mainly include point clouds and meshes. Euclidean data are usually of rasterized forms, they have regular grids. For example, images are composed of pixels which are well aligned and always have same number of neighbours.

Non-Euclidean data are usually more of geometric forms, they do not have regular grids. For example, with geometric metrics, the distance between two vertices on a mesh should be computed as their geodesic distance on the manifold, other than the direct Euclidean distance. In this section, different 3D data formats are briefly reviewed and compared.

## 2.1 Euclidean data

**Multi-view images:** 3D data may be presented as a combination of multiple 2D images captured for the 3D object from different view points [38]. Learning with this format, the noise effect from incompleteness, occlusion and illumination problems can be well reduced. All the input views jointly optimize the functions to represent the whole 3D shape. However, this format requires too many input sources and is usually too expensive for industrial use. The question of how many views are sufficient to represent a shape is also still open.

**RGB-D images:** With the development of RGB-D sensors, e.g., Microsoft Kinect, more and more industrial applications are using RGB-D images as the input data format for their tasks. This data format provides an additional depth map along with the normal 2D RGB color information. Comparing to other 3D data formats, there are more RGB-D data format available due to its inexpensiveness [7].

**Volumetric data:** Same as 2D shapes can be rasterized into pixels, 3D shapes can also be rasterized into voxels. In this case, 3D shapes are encoded by those occupied voxels. Despite the simplicity of the voxel-based representation, it suffers from keeping the intrinsic properties of 3D shapes and the smoothness of their surfaces [34]. It also requires high memory storage and has high computation complexity, which makes volumetric format not appropriate for high-resolution data.

## 2.2 Non-Euclidean data

**Point clouds:** A point cloud is a set of unstructured points that approximate the geometry of an object. However, if we only consider the local structure of

the object, those subsets may also be considered as Euclidean since they have a global parameterization and are usually represented by a normal system of coordinates. It depends on the metrics method that is used. But most tasks still focus on the global structure for shape recognition, matching or retrieval, hence point clouds are still classified as Non-Euclidean data format in most cases. Nowadays we have multiple choices of 3D sensors to generate point clouds, e.g., Ensenso or Zivid, they usually do single-shot and capture the whole scene. Therefore, different from other formats, preprocessing steps such as noise filtering or scene segmentation are usually required for point clouds of 3D shapes.

**Meshes/graphs:** A polygon mesh is a collection of vertices, edges and faces that defines the shape of a polyhedral object in 3D computer graphics and solid modeling. With an appropriate number of vertices, meshes can give extremely accurate geometric information of 3D shapes. The vertices in a mesh have certain connectivities, which makes mesh a special case of graph. The process of generating an approximate watertight mesh from a random connected graph is called 3D shape completition or inpainting. Although meshes contains rich information of 3D shapes, it is really a challenging task to learn on them directly due to its irregularity. In most relevant researches, the spectral properties of the graphs and meshes are utilized to learn latent features after applying a graph Laplacian eigen-decomposition.

**Continuous space function:** Continuous space functions are a very special data format. It uses a mathematical function to represent the 3D shape directly and precisely. It is also referred as level set or signed distance function (SDF) with minor definition modification. Input a coordinate in the defined space, a SDF outputs a value whose sign (positive or negative) denotes that this point is outside or inside the shape boundary. For example, if the output space of a SDF is defined between $[-1, 1]$, the whole function may be considered as a mapping function $f : \mathbb{R}^3 \rightarrow [-1, 1]$. If $0$ is defined as the cutoff boundary, then all the points whose coordinates yield an output between $[-1, 0]$ after the mapping means they are inside the object surface, and vice versa. However, only simple shapes like cube, heart, donuts or lemon can be easily denoted with a SDF. It is more often impossible to find such a function for a slightly complex shape. Thus this data format is less explored comparing to others.

**Table 2.1**: Property comparison of different 3D data formats

| | | Accuracy | Affability to NNs | | Geometric manipulability | |
|---|---|---|---|---|---|---|
| | | | input/output | computation | deform etc. | constrains |
| Euclidean | image-based | controversial | great | great | poor | poor |
| | voxel-based | poor | good | poor | controversial | poor |
| Non-Euclidean | point clouds | good | good | controversial | good | good |
| | meshes/graphs | great | poor | controversial | great | good |
| | continuous space function | great | poor | poor | controversial | poor |

## 2.3 Property comparison

It is impossible to say which data format is the best 3D data format. Apart from the accuracy requirements, to better make use of the 3D information, it is usually excepted that the data should be geometrically manipulable (deformation, interpolation, etc.) and convenient to impose structural constraints. On the other hand, since we are interested in applying deep learning algorithms on them, the data should also be able to be easily formulated as the input/output to neural networks and make fast forward/backward propagation computation possible. Here, based on the state-of-the-art researches, we summarize the overall rating subjectively on these properties of different 3D data formats in Table 2.1. In most cases, people will just use the most appropriate data format for their tasks according to the input source limitation, computation ability, and accuracy and robustness requirements.

# 3 Latent representations of 3D data

The process of acquiring latent representations from input data is essentially a mapping process. It maps the input data from its original data space to another latent space, which are usually lower dimensional. In statistics definition, latent representations (or, latent variables) are variables that are not directly observed but are rather inferred through a mathematical model from other variables that are directly observed and measured. Although multi-view images or volumetric data may be regarded as a special mapping method that maps the original geometric data into a lower dimensional space, those data representations are usually not considered as latent ones since we can still observe shape properties directly on them. Hence, in this report, we regard them as other kinds of data formats and not as latent representations.

Before the recent upsurge of deep learning, there were already many other classical mathematical methods that try to encode 3D data, mostly on 3D shapes. For 3D shapes, the latent representations of them are also called as shape descriptors. In this section, we first make a brief overview on those classical

methods and the shape descriptors they generated, then probe into the latent representations learned with neural networks.

## 3.1 Classical methods

Classical methods usually have very strong mathematical background, involving strict mathematical formulas and deductions. Therefore the encoding results from them are usually deterministic. There are numerous classical methods that try to learn latent representations from 3D data, whether on Euclidean formats or Non-Euclidean formats. Here we just summarize and list some most known ones that may be related or helpful to our future work.

**Ray-based sampling with spherical harmonics:** In order to characterize shapes of functions on a sphere by just a few parameters, spherical harmonics [9] were proposed as a suitable tool. The magnitudes of complex coefficients, which are obtained by applying the fast Fourier transform on the sphere to the samples, are regarded as vector components. Thus, the ray-based feature vector is represented in the spectral domain, where each vector component is formed by taking into account all original input.

**Laplacian spectral eigenvectors:** In addition to considering the connectivity of nodes and edges in a graph, mesh Laplacian operators take into account the geometry of a surface (e.g. the angles at the nodes). For a manifold triangle mesh, the Laplace-Beltrami operator is used to represent the intrinsic geometric structure. After applying the Laplacian eigen-decomposition, the original shape may be represented by its spectral eigenvectors, which makes mesh processing [24] and surface editing [25]possible.

**Heat kernel signature:** A heat kernel signature (HKS) is a shape descriptor obtained via spectral shape analysis methods and in use for deformable shape analysis. It is based on heat kernel, which is a fundamental solution to the heat equation [27]. For each point in the shape, HKS defines its feature vector representing the point's local and global geometric properties. HKS is one of the many recently introduced shape descriptors which are based on the LaplaceBeltrami operator associated with the shape. There are other relevant

shape descriptors including global point signature (GPS), biharmonic signature (BS), wave kernel signature (WKS).

**Skeleton-based 3D descriptor:** Skeletons derived from solid objects can be regarded as intuitive object descriptions. They are able to capture the most important information about the shape structure. Sundar et al. [28] presented a framework for skeletonization and 3D object retrieval. Skeleton-based 3D descriptor is widely used in animation and film industrial nowadays due to its ideal parameterized control on the shape joints.

**Primitive-based CAD model descriptor:** 3D shapes may be approximately assembled by composing simple volumetric primitives including cuboids, cylinders and spheres. The shapes from one category usually have similar primitive representations. Using this abstract representation, interpolation between the obtained latent representations may provide a consistent parsing across shapes in one certain category.

## 3.2 Neural learning-based methods

Comparing to the classical methods, neural learning-based methods are less deterministic since they have more stochastic calculations involved. The final parameters of a trained neural network may be slightly different even though all the settings are identical in multiple trainings.

Actually, the latent representations learned via neural networks are seldom of particular concern in most computer vision tasks, while they have always been implicitly used. A good example would be the bottleneck features in transfer learning. In transfer learning, we take a pre-trained model including network and weights, then remove the last few fully connected (FC) network and construct our own in place of it. When the training starts on the new data set, usually the original network parameters before the FC network are frozen and only the newly added FC network are trained. Here the input to the FC network is referred as bottleneck features. They represent the latent features learned from the last convolution layer in the network. Surely we can take the feature maps from any previous layer and name them as bottleneck features or latent representations, but in most cases we are more interested in a vector representation, thus a flattened

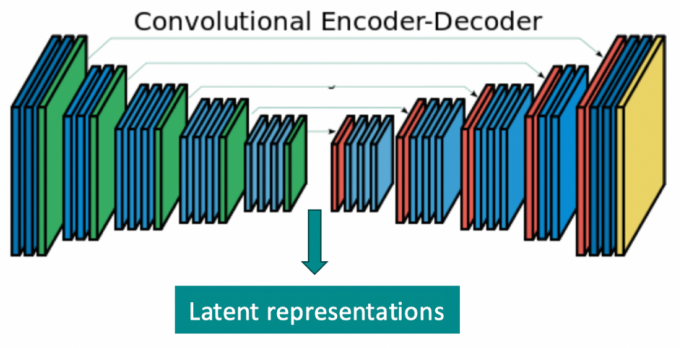



**Figure 3.1**: The basic structure of a neural encoder-decoder. The feature maps/vectors learned inside the network may be regarded as latent representations.

vector bottleneck feature are more often taken and used. But, still, the properties of bottleneck features themselves are really less explored.

Also in generation tasks, latent representations are also crucial to learning. A typical generative adversarial network (GAN) may take a vector from the latent space as the input to generate pseudo real world data. Interpolating between the input latent vectors, a continuous reshaping or deforming output can usually be observed.

Figure 3.1 gives a brief idea how latent representations are learned within a neural encoder-decoder. A more detailed survey of how latent representations of 3D shapes are obtained and utilized with deep learning methods is given in the next section.

# 4 Deep learning on 3D data

## 4.1 Learning on 3D Euclidean data

In order to duplicate the success of deep learning techniques from the 2D domain to the 3D domain, it is easy to see that we can use 3D Euclidean data directly

for learning purposes. In case of only 3D Non-Euclidean data are provided, we can always convert them into Euclidean formats with a certain information loss. Due to its simplicity and convenience, this converting process has been widely utilized to create rasterized data to fit in the Euclidean neural network architectures ever since the emerging of deep learning, even till now.

### 4.1.1 Image-based representations

When using RGB-D images or multi-view images as the input for deep learning tasks, it is often required to have multiple input channels or even multiple CNN streams to process the data. For example, [5] used a two-stream CNNs on RGB-D data for 3D object recognition tasks. The learned latent features from two streams were fused together in one later FC layer and the classification result was given after a further softmax layer. A more interesting method was proposed in [2], in which the idea of transfer learning was combined with the method used in [5]. It used four separate CNNs to train the four channels in the RGB-D data, while the weights were transferred from each network to another. Their results indicated that the depth information carries valuable information about shapes.

More processing streams will be needed for the multi-view images data format. MVCNN [26] processed rendered 12 views of a 3D object separately. Then a max pooling operation was applied in the view-pooling layer to get a compact latent representation for the whole shape. In [37], a multi-branch CNN has been designed to use rendered depth maps from different views of the object as input. Each branch returned a feature vector that contributes to the final classification. Apart from single value output recognition/classification tasks, this format has also been used for other more complex tasks. Kalogerakis et al. [11] designed a neural network for segmenting 3D objects into their labeled semantic parts by learning from their multiple 2D projections. Local shape descriptors from part correspondences have also been learned with a multi-view convolutional network [10]. Even 3D shape reconstruction via multi-view convolutional networks has also been studied from sketches in [13].

### 4.1.2 Volumetric data

Regular 2D convolution operations have been naturally extended to 3D convolution operations by applying 4D convolutional kernels, certain network architectures have also been proposed. VoxNet [15] first converted the point clouds of shapes into voxels according to their occupancy in the space. Then this volumetric data was used as input to their neural network for shape classification. A similar method has been propose in 3DShapeNets [33] except they got the volumetric data from depth maps. As a followed work, Seaghat et al. [23] modified the architecture of VoxNet by incorporating the orientation of 3D objects in the learning process.

Regarding the synthesis tasks with 3D volumetric data, in [32], by extending the idea of GAN in the 2D domain, volumetric generative adversarial networks have also been proposed. In McRecon network structure [8], foreground masks have been used as weak supervision through a raytrace pooling layer for 3D reconstruction. There are also octree-based methods which only consider the occupied grids in a more memory efficient way including OctNet[22] and O-CNN [29].

## 4.2 Learning directly on 3D Non-Euclidean data

As mentioned in the last subsection, people can always convert 3D Non-Euclidean data to Euclidean formats for convenient neural network architecture designs since the technical maturity of similar methods in 2D domain are already quite high. However, object information will be inevitably lost during the converting process. The best way to prevent this information loss is learning directly on 3D Non-Euclidean data, in which special ways to define the input, output, or even the operations used in the networks are usually required.

### 4.2.1 Point clouds

The very first proposed deep learning-based method of directly using 3D point clouds data for shape analysis tasks is PointNet [20]. It used $(x, y, z)$ coordinates of points as input to the network, then an additional spatial transform network

was performed as a pre-processing step. After that, lots of weights-sharing fully connected layers were added to compute point-wise features. Finally, a max-pooling layer was used to aggregate the global information and output a 1024 dimensional latent feature vector for classification tasks. For segmentation tasks, the global shape feature and the point-wise features were concatenated for predicting point-wise segmentation result. Despite the competitive results achieved by PointNet, it still failed to take full advantage of the local features in point clouds. Their subsequent work PointNet++ [21] tried to address this point by grouping the points with different scales, performing PointNet on them separately in order to aggregate different scale features. To better aggregate the information in the real local area, aggregate operations similar to the convolution operations have also been proposed, such as EdgeConv defined in [31] or X-Conv defined in [12]. Both of them took a certain number of neighbours of each point into consideration and performed the aggregating operation point-wise. With this operations, the learned final latent representation also contains local information implicitly.

In 3D point clouds synthesis field, [1] proposed a deep auto encoder (AE) with high reconstruction quality and generalization. Generative adversarial networks (GANs) and Gaussian Mixture Models (GMMs) have also been trained in the latent space of their AEs respectively. Similarly, FoldingNet [35] proposed a point clouds auto-encoder via deep grid deformation with graph-based encoders, in which special perceptron layers were defined as folding operations. Regarding the upsampling task for sparse point clouds, PU-Net was especially designed with convolution operations defined in the latent feature space [36].

### 4.2.2 Meshes

At first glance, triangular meshes give people the illusion that 2D convolutional kernels may be directly applied. However, these rasterized kernels are only applicable to Euclidean data due to their structure shift invariance property. In order to perform convolution locally, appropriate local patches need to be defined. Geodesic CNN (GCNN) [14] constructed local patches in local polar coordinates to ensure their structure non-position-dependent. Values of the functions around each vertex in the mesh are mapped into local polar coordinates

using the patch operator, thus geodesic convolution may be applied on those patches. Later on, Anisotropic CNN (ACNN) [3] was proposed to tackle the limitations in GCNN. It constructed a simpler pattern of local patches, which are independent to the injectivity radius of meshes. Rather than using a fixed kernel pattern as in GCNN and ACNN, MoNet [18] were proposed to define a vertex-wise locally weighted coordinate system, on which parametric kernels were applied to define the weighting functions. With this definition, GCNN and ACNN may be considered as special cases of MoNet with certain constraints.

Except for those methods defined on the spatial domain, methods defined on the spectral domain have also been proposed. For example, [6] first computed heat kernel descriptors of shapes based on their heat kernel signatures (HKS), then the descriptors were fed into two neural networks with target value using Eigen-shape Descriptor and Fisher-shape Descriptor, respectively. The final deep shape descriptor is formed by concatenating nodes in hidden layers. [30] proposed a similar pipeline with local point signature (LPS) features. Multi-scaled vertex spectral images were generated by packing the 16-dimensional LPS in a compact manner, and then fed into a CNN to generate the final shape descriptor. Those methods show the possibility that shape properties obtained via classical methods may be further utilized with the deep learning methods to get a better latent representation, with which better performance of different tasks may be achieved.

### 4.2.3 Continuous space function

Continuous space function (CSF) or signed distance function (SDF) is a really less explored data format. Although it provides high accuracy, it is usually impossible to easily find a function that matches a slightly complex object. Fortunately, neural networks are "universal approximators" and can mimic any continuous function to the degree that the network size permits.

Early this year, DeepSDF [19] was proposed to learn a continuous SDF representation for a 3D shape, which encoded a shape's boundary as the zero-level-set of the learned function that explicitly divided the space into shape interior and shape exterior. Deep Level Sets [17] also deployed a similar idea to represent the output as an oriented level set of a continuous embedding function with the

help of deep neural networks. In a more recent paper, Mescheder et al. [16] proposed Occupancy Networks, which also used a network to mimic functions that define the shape boundaries. An interesting adaption in their method is that rather than a signed value, the output of the network is a real value between 0 and 1, which indicates the occupancy possibility of a certain point in that space position. Although all those methods usually need a post-processing step to visualize the shapes, the reconstruction performance of them are usually qualitatively better than the performance of classical methods that only work for point clouds or meshes.

# 5 Conclusion

In this report, we first briefly review the most used 3D data formats, including both the Euclidean ones and the Non-Euclidean ones. Secondly, latent representations or shape descriptors obtained via classical methods and deep neural networks have been reviewed and discussed. While several classical methods have been addressed, more efforts have been put into investigating the neural learning-based methods. Latent representations of different 3D data formats learned with various network architectures have been reviewed and discussed, the possibility of combing classical methods and neural learning-based methods has also been especially addressed. Although within the deep learning scope, the dominant approaches that utilized for various computer vision tasks nowadays are still usually based on images or other Euclidean data, we hope that with a better learning and understanding of the latent representations of 3D shapes, more efficient architectures may be proposed and better performance may be achieved with them in the future.

# References

[1] P. Achlioptas et al. “Learning Representations and Generative Models for 3D Point Clouds”. In: *International Conference on Machine Learning (ICML)* (2018).

[2] L. Alexandre. “3D Object Recognition Using Convolutional Neural Networks with Transfer Learning between Input Channels”. In: *Intelligent Autonomous Systems* 13 (2016), pp. 889–898.

[3] D. Boscaini et al. “Learning Shape Correspondence with Anisotropic Convolutional Neural Networks”. In: *NIPS* (2016).

[4] B. Bustos and D. Keim and D. Saupe, T. Schreck, and D. Vrani. “Feature-based Similarity Search in 3D Object Databases”. In: *ACM Computing Surveys (CSUR)* 37 (2005), pp. 345–387.

[5] A. Eitel et al. “Multimodal Deep Learning for Robust RGB-D Object Recognition”. In: *Intelligent Robots and Systems (IROS)* (2015), pp. 681–687.

[6] Y. Fang et al. “3D Deep Shape Descriptor”. In: *CVPR* (2015), pp. 2319–2328.

[7] M. Firman. “RGBD Datasets: Past, Present and Future”. In: *CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis* (2016).

[8] J. Gwak et al. “Weakly supervised 3D Reconstruction with Adversarial Constraint”. In: *International Conference on 3D Vision (3DV)* (2017).

[9] D. Healy et al. “FFTs for the 2-sphere Improvements and Variations”. In: *Journal of Fourier Analysis and Applications* 4 (2003), pp. 341–385.

[10] H. Huang et al. “Learning Local Shape Descriptors from Part Correspondences With Multi-view Convolutional Networks”. In: *CVPR* (2017).

[11] E. Kalogerakis et al. “3D Shape Segmentation with Projective Convolutional Networks”. In: *CVPR* (2017).

[12] Y. Li et al. “PointCNN”. In: *arXiv preprint arXiv:1801.07791* (2018).

[13] Z. Lun et al. “3D Shape Reconstruction from Sketches via Multi-view Convolutional Networks”. In: *Proceedings of the International Conference on 3D Vision (3DV)* (2017).

[14] J. Masci et al. “Geodesic Convolutional Neural Networks on Riemannian Manifolds”. In: *IEEE International Conference on Computer Vision Workshop (ICCVW)* (2015).

[15] D. Maturana and S. Scherer. “VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition”. In: *International Conference on Intelligent Robots (IROS)* (2015), pp. 922–928.

[16] L. Mescheder et al. “Occupancy Networks: Learning 3D Reconstruction in Function Space”. In: 2019.

[17] M. Michalkiewicz et al. “Deep Level Sets: Implicit Surface Representations for 3D Shape Inference”. In: *ArXiv* abs/1901.06802 (2019).

[18] F. Monti et al. “Geometric Deep Learning on Graphs and Manifolds using Mixture Model CNNs”. In: *CVPR* (2017).

[19] J. Park et al. “DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation”. In: *CVPR* (2019).

[20] C. Qi et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *CVPR* (2017).

[21] C. Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *NIPS* (2017).

[22] G. Riegler, A. Ulusoy, and A. Geiger. “OctNet: Learning Deep 3D Representations at High Resolutions”. In: *CVPR* (2017).

[23] N. Sedaghat et al. “Orientation-boosted voxel nets for 3D object recognition”. In: *British Machine Vision Conference (BMVC)* (2017).

[24] O. Sorkine. “Laplacian Mesh Processing”. In: *Eurographics 2005 - State of the Art Reports* (2005).

[25] O. Sorkine et al. “Laplacian Surface Editing”. In: *Symposium on Geometry Processing* (2004).

[26] H. Su et al. “Multi-view Convolutional Neural Networks for 3D Shape Recognition”. In: *ICCV* (2015), pp. 945–953.

[27] J. Sun, M. Ovsjanikov, and L. Guibas. “A Concise and Provably Informative Multi-Scale Signature-Based on Heat Diffusion”. In: *Computer Graphics Forum* 28 (2009), pp. 1383–1392.

[28] H. Sundar et al. “Skeleton-based Shape Matching and Retrieval”. In: *Proceedings of the Shape Modeling International* (2003), pp. 45–53.

[29] P. Wang et al. “O-CNN: Octree-based Convolutional Neural Networks for 3D Shape Analysis”. In: *SIGGRAPH* (2017).

[30] Y. Wang et al. “A Robust Local Spectral Descriptor for Matching Non-Rigid Shapes with Incompatible Shape Structures”. In: *CVPR*. 2019.

[31] Y. Wang et al. “Dynamic Graph CNN for Learning on Point Clouds”. In: *arXiv preprint arXiv:1801.07829* (2018).

[32] J. Wu et al. “Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling”. In: *NIPS* (2016).

[33] Z. Wu et al. “3D ShapeNets: A Deep Representation for Volumetric Shapes”. In: *CVPR* (2015).

[34] Y. Xiang et al. “Data-driven 3D Voxel Patterns for object category recognition”. In: *CVPR* (2015), pp. 1903–1911.

[35] Y. Yang et al. “FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation”. In: *CVPR* (2018).

[36] L. Yu et al. “PU-Net: Point Cloud Upsampling Network”. In: *CVPR* (2018).

[37] P. Zanuttigh and L. Minto. “Deep learning for 3D shape classification from multiple depth maps”. In: *International Conference on Image Processing (ICIP)* 155 (2017), pp. 3615–3619.

[38] S. Zhi et al. “Toward Real-time 3D Object Recognition: A Lightweight Volumetric CNN Framework Using Multitask Learning”. In: *Computers and Graphics* 71 (2017), pp. 199–207.

## Karlsruher Schriftenreihe zur Anthropomatik (ISSN 1863-6489)

Herausgeber: Prof. Dr.-Ing. habil. Jürgen Beyerer

**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

Lehrstuhl für Interaktive Echtzeitsysteme
Karlsruher Institut für Technologie

Fraunhofer-Institut für Optronik, Systemtechnik
und Bildauswertung IOSB Karlsruhe

In 2019, 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) has again been hosted on the Griesgeth- of nearby the town of Triberg-Nussbach in Germany. For a week from July, 29 to August, 2 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
ISBN 978-3-7315-1028-4