Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 847

Gregor Blott

Multi-View Person Re-Identification

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5259-8

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover
ISSN 0174-1454, Nr. 356, Hannover 2020

Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 847

Multi-View Person Re-Identification

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation

Vorgelegt von

Gregor Blott, M. Sc.

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5259-8

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover
ISSN 0174-1454, Nr. 356, Hannover 2020

Adresse der DGK:

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften (DGK)
Alfons-Goppel-Straße 11 ● D – 80 539 München
Telefon +49 – 331 – 288 1685 ● Telefax +49 – 331 – 288 1759
E-Mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission:

Vorsitzender:	Prof. Dr.-Ing. Steffen Schön
Referent:	Prof. Dr.-Ing. Christian Heipke
Korreferenten:	Prof. Dr.-Ing. Claus Brenner Prof. Dr.-Ing. Jürgen Gall

Tag der mündlichen Prüfung: 15.04.2020

ISSN 0065-5325 ISBN 978-3-7696-5259-8

Abstract

Appearance-based person re-identification is a crucial component of multi-camera networks in different applications such as surveillance, automation, and business analytics.

Despite considerable recent progress, the performance is still not satisfactory for autonomous deployment in practice. This is due to the high intra-person variation, the low inter-person variation, and the significant complexity of the task. Generally speaking, the issues are that the necessary number and variety of observations to uniquely describe and differentiate among persons, are typically not available.

We present a novel approach for appearance-based person re-identification, which exploits within-video multi-view information of fisheye cameras looking downwards from the ceiling. To handle these highly variable multi-view observations, we develop a generic pipeline for processing fisheye camera imagery based on prior knowledge, geometric sensor modeling and deep learning.

In an extensive experimental evaluation, we show that: i) Fisheye cameras are a useful tool to obtain multiple views of persons from single cameras. ii) Multi-view observations improve the inter-camera re-identification of persons by a large margin. This indicates that our novel system-level development boosts the performance, mainly due to much more variety in observations. iii) The boost is irrespective of using more training data and the applied feature extraction method.

As a consequence of our work, a novel approach is available, which allows for inter-camera person re-identification using multiple person views per camera by explicitly exploiting fisheye distortions.

Keywords: *person re-identification, within-video multi-view, fisheye lens, deep learning*

Zusammenfassung

Erscheinungsbasierte Personenwiedererkennung ist eine entscheidende Komponente für unterschiedliche Anwendungen von Kameranetzwerken wie beispielsweise der Sicherheitskamera- und der Automatisierungstechnik sowie der Analyse von Kaufverhalten in Geschäften.

Trotz großer Fortschritte in den letzten Jahren, ist die Wiedererkennungsleistung für praktische Anwendungen aufgrund der hohen Variabilität des Aussehens von Personen und einer hohen Szenenkomplexität noch nicht ausreichend gelöst. Das grundlegende Problem kann auch als ungenügende Beobachtungen und eine geringe Variabilität von Beobachtungen bei klassischen Kameraansätzen aufgefasst werden, wodurch Personen nicht eindeutig wiedererkannt werden können und viele Mehrdeutigkeiten entstehen.

In der Arbeit wird ein neues Verfahren für die Personenwiedererkennung vorgestellt, das Bilder von Fischaugenkameras verwendet, die an der Decke montiert sind und in Nadirichtung schauen, um Personen aus unterschiedlichen Ansichten (Multi-View) zu analysieren. Um die verschiedenen Ansichten zu prozessieren, wird auf Basis von domänenspezifischem Vorwissen, geometrischer Sensormodellierung und Deep Learning eine generische Verarbeitungskette präsentiert.

Experimente bestätigen: i) Fischaugenkameras in Nadirrichtung sind in der Lage Bilder von Personen aus unterschiedlichen Blickrichtungen pro Kamera bereitzustellen. ii) Personenwiedererkennung mit mehreren Ansichten verbessert datensatzspezifisch die Leistungsfähigkeit enorm. iii) Darüber hinaus ist der Ansatz unabhängig von mehr Trainingsdaten oder einer anderen Merkmals-Extraktion-Methode, da wir bewusst diese als generisches und austauschbares Modul ausgeführt haben.

Die Konsequenz dieser Arbeit ist, dass ein neuartiges Verfahren zur Verfügung steht, um Personen aus unterschiedlichen Ansichten zu beobachten und bei gegebenen Beobachtungen aus anderen Kameras, wiederzuerkennen. Hierbei werden bewusst Fischaugenverzeichnungen ausgenutzt.

Acknowledgments

This work is supervised by Prof. Dr.-Ing. Christian Heipke, chair of the Institute of Photogrammetry and GeoInformation (IPI), Leibniz Universität Hannover (LUH). I want to thank him for offering me the opportunity to investigate multiple aspects of photogrammetry and computer vision in the context of my research and beyond, providing a reliable research environment, along with a fruitful exchange with other international scientists.

Moreover, I want to thank my thesis reviewers, Prof. Dr.-Ing. Claus Brenner and Prof. Dr.-Ing. Jürgen Gall, for their very pertinent advice and remarks, as well as the thesis committee president for making his time available for me.

My gratitude also goes to the executives and staff of the Computer Vision Research Lab (CV Lab) of the Robert Bosch GmbH, Hildesheim. The CV Lab funded this research and allowed me to pursue academic research after having contributed to industrial research for some years.

In addition, I would like to thank colleagues of both affiliations, the IPI and the CV Lab, for fruitful discussions. Many thanks especially to Jan Rexilius for his input, excellent questions and remarks at any stage and any time of my work, Jie Yu for regular meetings and exchange, Prof. Dr. techn. Franz Rottensteiner for various valuable discussions, Max Mehltretter and the members of the research training group i.c.sens for our weekly exchange while sharing an office.

Further thanks go to my parents for their understanding and support.

Finally, I thank Elisabeth and Cleonie for encouragement throughout eventful times.

Contents

Definition, Acronyms and Symbols

Definition

As a "**view**", we define the viewing-direction in which a person is seen. For example, a person is observed from the front, from the back, from the side, or from a bird's eye view. In order to avoid misunderstandings, we point out that other authors refer to the same with the following terms: "viewpoint", "view information", "viewing-direction", "view-angle", "rotation-angle", and "viewing perspective".

Acronyms

BEV	Bird's eye view
FE	Fisheye
IDs	Identifiers
PRID	Person Re-Identification
rank#1	top#1 rank of the CMC curve
re-id	Re-Identification

General notation

$[a, b, \alpha, \beta]$	Scalars
$[\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{X}, \mathbf{\Phi}, \Delta]$	Vectors
$[\mathbf{A}, \mathbf{B}, \mathbf{K}, \Sigma]$	Matrices

Projection

$\mathbf{P} = [X, Y, Z]^T$	3D world point in Cartesian coordinates
$\mathbf{p} = [x, y]^T$	2D image point
$\mathbf{K}$	Calibration matrix
c	Calibrated focal length
$[x_0, y_0]^T$	Principal point

Person Signature

General notation

$\mathcal{P}$	Unknown person (probe)
$\mathcal{G}_i$	Known person (from gallery) with id i
$\mathcal{G} = \{\mathcal{G}_1, ..., \mathcal{G}_n\}$	Gallery of n known persons

Image space

$I_\mathcal{P}$	Image of unknown person (probe)
$I_{\mathcal{G}_i}$	Image of known person (from gallery) with id i
$I_\mathcal{G} = \{I_{\mathcal{G}_1}, ..., I_{\mathcal{G}_n}\}$	Images of known persons from gallery
$I \in \{I_\mathcal{P}, I_\mathcal{G}\}$	Any image consisting of a person

Latent space

$\Phi_\mathcal{P}$	Embedding in latent space of unknown person (probe)
$\Phi_{\mathcal{G}_i}$	Embedding in latent space of known person (from gallery) with id i
$\Phi_\mathcal{G} = \{\Phi_{\mathcal{G}_1}, ..., \Phi_{\mathcal{G}_n}\}$	Embeddings in latent space of known persons from gallery
$\Phi \in \{\Phi_\mathcal{P}, \Phi_\mathcal{G}\}$	Any embedding consisting of a person

Chapter 1

Introduction

This thesis proposes a novel approach for person re-identification with multi-view observations. In this chapter, we discuss the general motivation, present the problem statement, and give an overview of the thesis and its contributions.

1.1 Motivation

Camera systems are gaining in popularity in several daily life cases, including smartphones, the automotive context, robotics, consumer goods, and various others. Using more than one camera with spatial separation and communication capabilities spawns a camera network, sometimes also called multi-camera network [Aghajan & Cavallaro, 2009], which enables novel applications compared to a single camera.

A security camera network consists of security cameras, also known as surveillance cameras, Closed Circuit Television (CCTV) cameras or forensic cameras. It is applied for diverse use-cases to pro-actively or forensically investigate scenarios in private and public domains such as fairgrounds, railway stations, airports, subway stations, shopping complexes, supermarkets, campuses, and others (cf. Fig. 1.1). Due to historical and recent threats a demand for more security exists, which is evident from a rapidly growing number of camera installations all over the world and an increasing market size [Statista.com, 2018].

Already in 2013 the British Security Industry Authority estimated that up to 4.9 million security cameras exist in the country, including 750,000 in sensitive locations such as schools, hospitals, and care homes [Telegraph, 2013]. As storage costs have decreased by using cloud-based solutions and more data are stored than in the past, it is a tremendous task for a human operator to analyze all the hours of video material. Therefore, automatic image interpretation is becoming more and more essential.

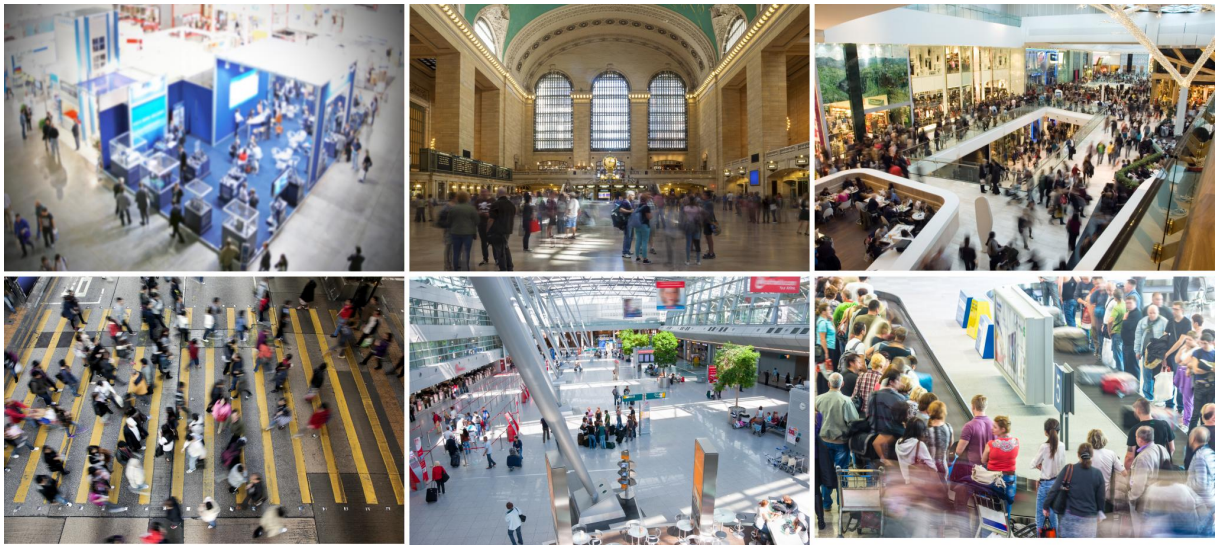


Figure 1.1: Typical scenarios where camera networks are applied. © Images licensed by www.depositphotos.com

Also in other domains such as robotics [Aghajan & Cavallaro, 2009], automotive [Schön et al., 2018], and business analytics [Hakeem et al., 2012], camera networks play an increasing role in establishing new business models or improving the reliability of results. Possible applications are, for example: i) the analysis of global person trajectories in stores to find poorly visited shelves, ii) the retrieval of previously seen persons from another disjoint camera field of view, e.g. to find a missing child in a fair complex, iii) in terms of smart homes, the re-identification of individuals to setup preferred environmental conditions such as lighting, heating, and others.

In this work, we focus on a mandatory prerequisite for various camera network applications, the automatic re-identification of previously seen individuals. This challenging task is called *person re-identification*.

1.2 Person Re-Identification

The term *Person Re-Identification* (PRID) is not used unambiguously and is sometimes misinterpreted. One of the oldest descriptions of *re-identification* is given in [Plantinga, 1961], where the relationship of mental states and behavior is discussed as *"to re-identify a particular, then, is to identify it as (numerically) the same particular as one encountered on a previous occasion."*

Another description of PRID is given in [Zajdel et al., 2005] and names the task of re-identifying a person after he or she left the field of view and re-entered later. In [Zajdel et al., 2005] spatial-temporal and visual cues are used to distinguish between persons. However, by using the visual cues, which change after persons modify their appearance, "later" is inherently limited to short-term re-identification.

In the past, different organizations defined terms such as detection and recognition, for example in the NATO Standardization Agreement [STANAG, 1998] or in the international engineer

standards such as IEC 62676 [IEC, 2014]. According to STANAG a definition of detection is: *"In imagery interpretation the discovering of the existence of an object, but without recognition of the object"*, the definition of recognition is: *"The ability to fix the identity of a feature or object in imagery within a group type"*. However, a definition of re-identification was not given. In 2011, the European Commission published a so-called *Orientation Material* (EUROSUR-2011) [Frontex, 2011], it says :

> **Detect**:
> *To establish the presence of an object and its geographic location, but not necessarily its nature.*
> **Classify**:
> *To establish the type (class) of object (car, van, trailer, cargo ship, tanker, fishing boat, ...).*
> **Identify**:
> *To establish the unique identity of the object (name, number), as a rule without prior knowledge.*
> **Recognise**:
> *To establish that a detected object is a specific pre-defined unique object.*
> **Verify**:
> *Given prior knowledge on the object, can its presence/position be confirmed.*

In recent years different publications have followed the orientation material e.g. [Gong et al., 2014, Nambiar, 2017], even though re-identification was not defined. However, Gong et al. agreed with [Frontex, 2011] and place person re-identification in the middle between identification and recognition, *"whenever a specific query with a target person is provided, and all the corresponding instances are searched in the dataset"* [Gong et al., 2014, p. 344].

To conclude, the term person re-identification is not linked by definition to any use-case, time lag, nor modality; rather a task is described, namely that a previously seen object, viz. a person, is re-identified as the same particular. Therefore, modalities such as soft biometrics[1] and hard biometrics[2] (cf. [Nambiar, 2017]) as well as appearance information or tracking strategies[3] could be used to solve the task. All modalities have pros and cons such as short-range, far-field or mandatory auxiliary assumptions[4] resulting in higher or lower re-identification accuracy due to an increasing number of ambiguities.

We understand person re-identification as the image-based task where individuals are initially observed by one camera-node of a multi-camera network to build a gallery of known candidates. Afterwards, persons observed in another camera-node are re-identified as the same particular person as in the gallery. The camera-nodes comprise of a disjoint field of view. Finally, a short-term re-identification, which terminates when a person changes their appearance, is our objective.

[1] Soft biometrics: e.g. eye color, body measurements, hair / beard / mustache and gait
[2] Hard biometrics: e.g. fingerprint, face, DNA, iris / retina, and palm print
[3] Tracking strategies: e.g. position in 3D space with respect to floor plan, or Wi-Fi beacons
[4] Auxiliary assumptions: e.g. WiFi, cooperative persons, particular camera pose

Figure 1.2: Challenges of appearance-based person re-identification. Left: high intra-person variation, one and the same person is shown. Right: low inter-person variation, four different persons are shown who share a very similar appearance.

1.3 Problem statement and research objective

To allow PRID in camera networks, we follow an appearance-based strategy. We motivate the choice to use purely appearance-based-information by the fact that it does not need any collaboration of persons or particular camera poses. Even if biometrics[1,2] are hard to fake [Nambiar, 2017] and therefore a benefit, especially for long-term re-identification, cooperative persons, particular camera poses, and controlled environments are a mandatory prerequisite to observe, e.g., the face, the gait or the iris. To name an example, face recognition was shown in the past to be a far-range high accuracy strategy (cf. the leaderboard of [Guo et al., 2016]). However, without particular camera orientations or with uncooperative persons, e.g. turning their faces away [Gong et al., 2014], it is a tremendous task to obtain good results. Nonetheless, we point out that the appearance-based re-identification is inherently limited to the point where the persons change clothes. This can be seen as a substantial benefit in terms of privacy since the association to a unique ID - e.g. passport ID - is hardly possbile with this single modality. However, for several applications, appearance can completely fulfill the requirements.

Appearance-based PRID is a rapidly growing research topic; see also the literature review in Section 2. The major challenge of person re-identification is the high intra-person variation and low inter-person variation due to a typically limited amount and variety of detailed observations in image space.

Intra-person variation can be defined as follows: one and the same individual appears in a first camera view, e.g. observed in a front view, differently than in another camera, where the person is, e.g., observed from the back. Therefore, the intra-class variation of the same person is very challenging for the re-identification. One example is given in the left part of Figure 1.2.[5]

Inter-person variation adapted for PRID can be explained as follows: given one unknown person in one camera, other persons across different cameras share a very similar appearance and look like being the correct corresponding match. One example is different men in black suits. Furthermore, persons in uniforms are hard to distinguish (cf. also right part of Figure 1.2).

[5] Much more challenging samples are provided in Fig. 4.1.

Against the background of the issues for person re-identification concerning reliability and accuracy, the research objective of this dissertation can be stated as follows:

This work aims at the improvement of appearance-based person re-identification for the security camera domain, by introducing multi-view information and integrating it as strong prior knowledge in the re-identification process. Additionally, we elaborate an approach to obtain purely image-based multi-view information with a single camera.

Note, detection and tracking of persons are not the focus in this thesis; the respective annotations are assumed to be available and are used as input for our objective.

1.4 Contribution

In this section the five contributions of this thesis are presented:

① We contribute a novel multi-view person re-identification approach, that consists of i) a fisheye camera setup to systematically obtain different viewing-directions of persons in every single camera of a camera network, and ii) a multi-view processing strategy to tackle the issues which stem from high intra-person and low inter-person variation. To the best of our knowledge, it is the first attempt to improve appearance-based person re-identification by proposing a monocular-camera based setup, which is inherently designed to increase the number and variety of views of persons.

② We develop and evaluate different methods for view-aware data fusion to find the best method. Afterwards, the best method is applied for the multi-view experiments on real fisheye data. To the best of our knowledge, we conduct the first person re-identification experiments with fisheye images and multi-view purpose.

③ We develop and evaluate a method to classify views of persons from single images by adapting a cross-domain classification method.

④ We conduct the first study of appearance-based person re-identification with *bird's eye views*. These views are provided by our proposed fisheye camera setup whenever persons pass the central field of view.

⑤ Due to the lack of existing datasets, we designed, recorded, re-mapped, and annotated **11** different datasets to conduct the experiments for the research underlying this thesis. Eight of these datasets were created from scratch.

Parts of this thesis have been published in a journal article as well as in peer-reviewed conference and workshop proceedings. The content of the following chapters is partly adapted from our works [Blott et al., 2019], [Blott et al., 2018b], [Blott et al., 2018a], [Blott & Heipke, 2017].

1.5 Outline of this thesis

The content of this thesis is structured as follows: **Chapter 2** provides a comprehensive overview of related work including the introduction of the commonly used terminology and strategies. In **Chapter 3** the fundamentals of this thesis are presented. These are, i) a fisheye lens and the corresponding geometric projection model, and ii) feature extraction methods which are used as a module in our approach. **Chapter 4** proposes our approach, and justifies the key-design decisions, along with a discussion on expected strengths and weaknesses. Furthermore, in **Chapter 5** the results of conducted experiments are presented and discussed. Finally, in **Chapter 6** conclusions are drawn and future research directions identified.

Chapter 2

Related work

In this chapter, we discuss work which is related to our research objective. Before starting with a comprehensive overview of current literature, a general introduction and brief historical overview are given, and additionally, the terminology is introduced.

2.1 Scope

Person Re-Identification (PRID) in general is a hot research topic [Zheng et al., 2016b], which is demonstrated by the significantly increasing number of publications per year (cf. Figure 2.1). As we follow an appearance-based re-identification strategy, we reduce the related work to this particular research field. Therefore, related work dealing, for example, with the following topics are beyond the chapter's scope: i) Non-video sensors, such as the Microsoft Kinect™ [Imani & Soltanizadeh, 2016] to use *anthropometric measures*. ii) A *passive bifocal stereo camera system* [Blott & Heipke, 2017] to extend the range and more importantly, to reconstruct spatial 3D information. iii) Taking the *camera topology*[1] into account [Cho et al., 2017].

2.2 Historical overview

Following [Zheng et al., 2016b], in 1997, a first work [Huang & Russell, 1997] addressed PRID for multi-camera tracking in a camera network without formally using the name "PRID".

In 2005, spatial and temporal information was used and the first definition of PRID was presented [Zajdel et al., 2005] (cf. Section 1.2).

In 2006, PRID was separated from multi-camera tracking and became an independent computer vision task [Gheissari et al., 2006]. The detection and tracking of persons were separated

[1] This is person re-identification considering the spatial relation between different cameras in a global context.

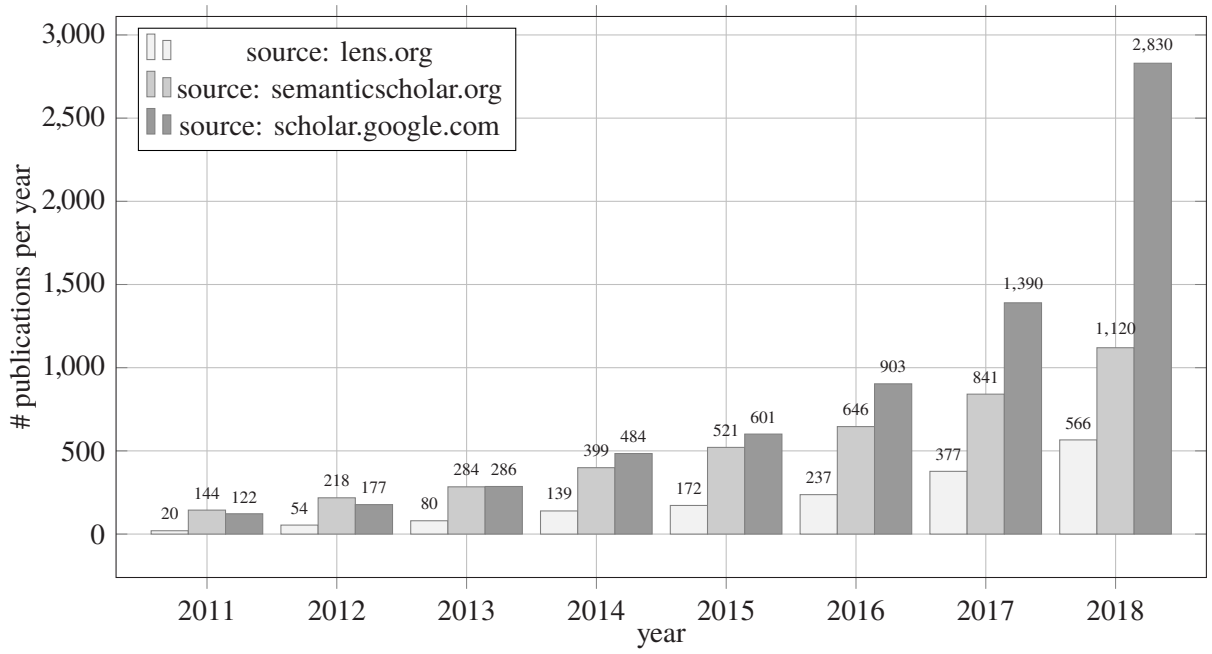


Figure 2.1: Publications dealing with the keyword: "*person re-identification*" (including conference proceedings, journals, arxiv and technical reports). The number of publications per annum is rapidly increasing. Database request November/30/2019.

from the re-identification part and the latter was addressed as an independent research topic. Consequently, cropped images of persons have been used as input for the re-identification. This is still the case today. Only few PRID publications focus on the complete pipeline, including detection, tracking, and selecting the best samples.

In 2010, [Bazzani et al., 2010, Farenzena et al., 2010] proposed *multi-shot* PRID. In contrast to single-shot, multiple images per camera were selected from a video sequence to describe a person. It was shown that multi-shot appearance information in combination with segmentation and multiple-frames not surprisingly outperforms a single-shot strategy. However, [Bazzani et al., 2010] also showed that the re-identification accuracy saturated as the number of selected frames was increased.

In 2014, the exploration of *Deep Learning* for PRID started when [Yi et al., 2014, Li et al., 2014] employed a *Siamese Neural Network* [Bromley et al., 1993] to determine if a person image input pair belongs to the same person. Even if the performance of handcrafted features in combination with metric learning still outperformed deep learning methods in those days, a rapidly growing community was established.

2.3 Terminology and strategies

Person re-identification can be expressed as follows: First, q persons are observed in different cameras, typically of disjoint fields of view. These q persons build the so-called gallery

$$\mathcal{G} = \{\mathcal{G}_1, ..., \mathcal{G}_o, ...\mathcal{G}_q\}, \tag{2.1}$$

whereas an imaged person in the gallery can be expressed as

$$\mathcal{G}_\mu = \{I_{\mathcal{G}_\mu,t}\}_{t=1,T_{\mathcal{G}_\mu}}, \tag{2.2}$$

and $\{I_{\mathcal{G}_\mu,t}\}_{t=1,T_{\mathcal{G}_\mu}}$ are the $T_{\mathcal{G}_\mu}$ images of the person with Id μ.

Then, an unknown person, called the "probe" ($\mathcal{P}$), is observed within another camera field of view in $T_\mathcal{P}$ images (I),

$$\mathcal{P} = \{I_{\mathcal{P},t}\}_{t=1,T_\mathcal{P}}. \tag{2.3}$$

Single-shot person re-identification is the strategy where an image pair is matched ($T_\mathcal{P} = T_{\mathcal{G}_\mu} = 1$). Matching between two sets of images is known as *multi-shot* ($T_\mathcal{P}, T_{\mathcal{G}_\mu} > 1$). Continuous time frames (videos) or randomly sampled images from continuous time frames are typically employed.

Closed-set person re-identification describes a process where a probe re-appears in the gallery.

$$ID(\mathcal{P}) \in \{ID(\mathcal{G}_1), ..., ID(\mathcal{G}_o), ...ID(\mathcal{G}_q)\}, \tag{2.4}$$

where $ID(\bullet)$ is a function to obtain the identity of $\mathcal{G}_o$, so that $ID(\mathcal{G}_o) \Rightarrow o$ and $ID(\mathcal{P}) \Rightarrow o$.

In this case the re-identification is modelled as a closed-world problem. This strategy is used for most PRID approaches [Zheng et al., 2016b] and formulates PRID as a retrieval problem where an ordered list of possible candidates with decreasing similarity can be used.

Then, given a probe ($\mathcal{P}$), the identity is determined by:

$$i^* = argmax_{i \in (1,...,\mu,...q)} sim(\mathcal{P}, \mathcal{G}_i), \tag{2.5}$$

where i^* is the best match and $sim(\cdot,\cdot)$ is a similarity measure which typically transforms the images into a feature space before the similarity is determined.

On the contrary, in *open-set* re-identification persons can appear only in the probe camera(s) or only in the gallery camera(s) and re-identification approaches additionally have to decide if an unknown person is present in the gallery or was never seen before [Zheng et al., 2016b]. The *closed-set* can also be seen as a special case of the *open-set* person re-identification. In practice

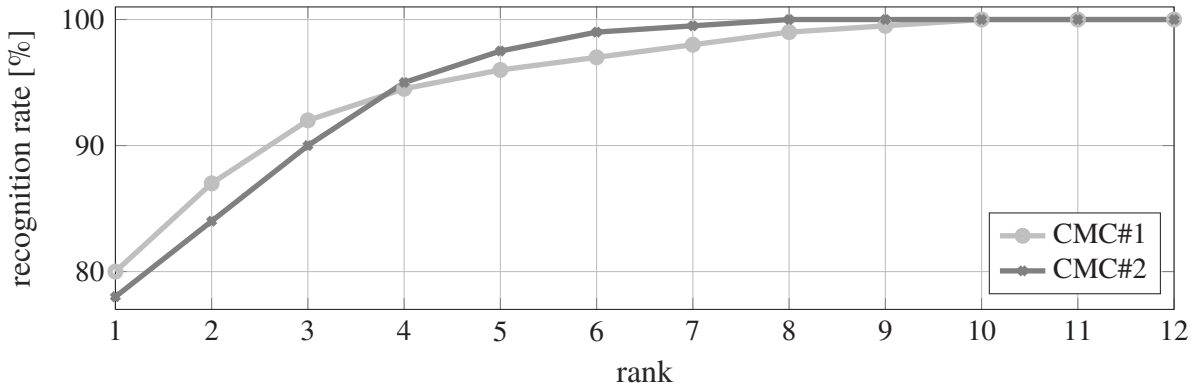


Figure 2.2: Two toy-examples of CMC curves. CMC#1 shows that the rank#1 recognition rate is 80%. This means that the best match, with 80% probability, is the correct match. When considering the rank#4, i.e. the correct match is between rank#1 and rank#4, the recognition rate is 94.5%. CMC#2 has a lower rank#1 recognition rate but reaches 100% recognition rate earlier. Thus, when rank#8 is considered the approach of CMC#2 is better. However, if the goal is that the best match is the correct match, the approach of CMC#1 shows the best performance.

closed-set person re-identification only works for restricted areas (e.g. shops) where persons use entry and exit doors and are directly observable by the cameras. Whenever a person enters or leaves the area from an unobserved space, the re-identification might fail. Note, for the *open-set*, the term PRID is not precise, since an unknown person who is not included in the gallery cannot be re-identified as the same person, the person is instead only identified as an unknown person. Here the term verification is a better fit.

The standard PRID *Evaluation Metric* for *single ground truth*[2] , the *closed-set*, and a pair of two cameras, one to build a gallery and the other to obtain the probes, is the cumulative matching characteristics (CMC) curve [Gray et al., 2007], see also Figure 2.2. Here, the re-identification is considered as a ranking problem. In particular, a CMC curve is generated after a test dataset was evaluated, and it shows the recognition rate as a function of the considered matching ranks. The recognition rate can be calculated as $\frac{n_c \cdot 100}{n_T}$, whereas n_c is the number of correctly re-identified images, and n_T the total number of images. Depending on the rank#r, an image is correctly re-identified if it is under the best r-ranks. In other words, the CMC curve shows the proportion of actual positives that are correctly re-identified as such, also the probability that a correct match appears within the top-k ranks is shown.

CMC is a valid measure if the focus is on returning the correct match within the top-k ranks. However, for datasets or applications where multiple ground truths exist[3], this metric only evaluates the top match of a person. If several ground truths exist, i.e. when multiple images per person are available in the gallery, the best of the multiple images is evaluated. Thus, if a gallery consists of images from multiple cameras, the probe is only found in one camera, where the person shows the highest similarity. With the rise of larger datasets to efficiently apply *Deep Learning* approaches, [Zheng et al., 2015a] proposed *mean average precision* (mAP) to model

[2] This is the case, if only one corresponding image exists in the gallery for a given probe ($T_{\mathcal{P}} = T_{\mathcal{G}_\mu} = 1$).
[3] $T_{\mathcal{P}}, T_{\mathcal{G}_\mu} > 1$ or persons are observed by multiple cameras.

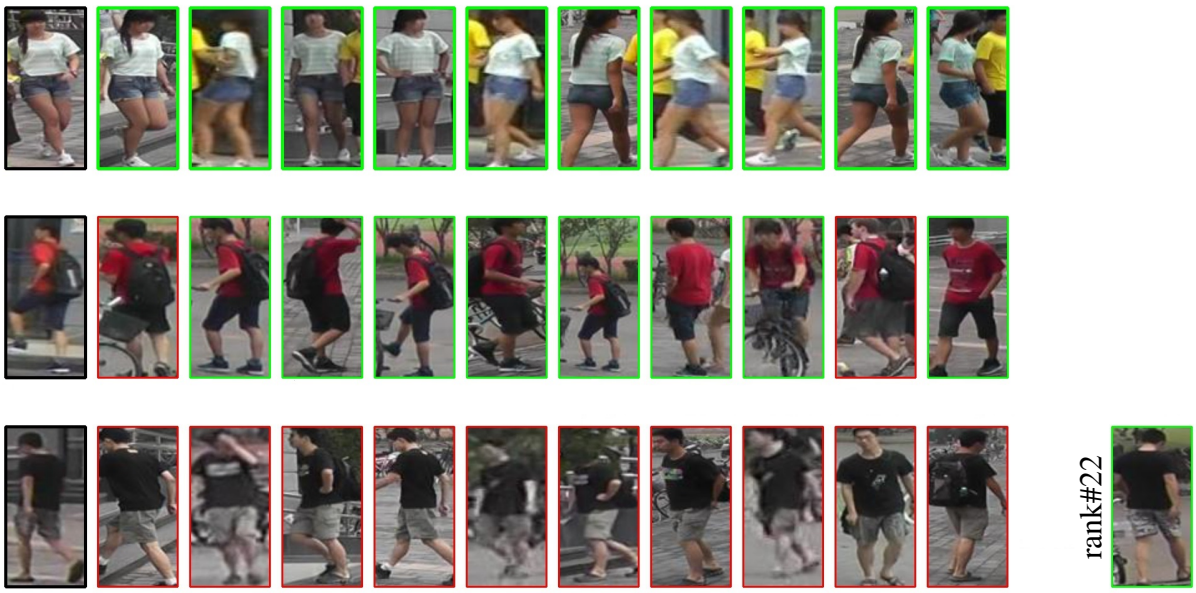


Figure 2.3: PRID as retrieval task: three unknown persons (probes) are requested (one per row). Probes (left), are bordered by black bounding boxes. From left to right, the ten most probable matches are depicted (rank#1 - rank#10) with decreasing similarity. Correct matches are displayed in green bounding boxes, false matches in red bounding boxes. The first row shows a very good result. All the top ten matches are correct. The second row shows an example where the first and ninth match are wrong. The third row shows an example where all ten matches are wrong, and the correct match is on rank#22. Cropped images of persons are taken from the Market-1501 dataset [Zheng et al., 2015a].

accuracy with respect to multiple-ground truths in the dataset. mAP takes recall and precision into account and is used for most datasets today.[4]

However, for datasets where only one image per candidate exists in the gallery, rank#1 is still the key performance indicator. Illustrative examples of person re-id visualization in general are given in Figure 2.3.

In this work, we use the rank#1 of the CMC curve as measure, consequently merely the matches with the highest similarity to a given probe are evaluated. We decided to employ this measure since we are only interested in finding the correct match, which is essential for an autonomous multi-camera-system. Furthermore, in our experimental evaluation, matching is only employed pairwise between two cameras of a multi-camera-network. Consequently, mAP and rank#1 are the same for our experiments because only one correct match exists in the gallery, and only the best match is evaluated.

A term which is not consistently used in the literature but needs to be defined in the course of this work, is *view* to emphasize the direction from which a person is seen, i.e., from the front, from the back, from the side or finer quantizations. In order to avoid misunderstandings, we point out that other authors name the same idea with the following terms: "viewpoint", "view information", "viewing-direction", "view-angle", "rotation-angle". Additionally, some authors

[4] "*For each query, the area under the Precision-Recall curve is calculated, which is known as average precision (AP). Then, the mean value of APs of all queries, i.e., mAP, is calculated, which considers both precision and recall of an algorithm.*" [Zheng et al., 2015a, p. 1119]

use a different meaning for "view", e.g., the influence of different camera orientations. In this thesis, the term *view* is employed as the term to describe the viewing-direction from which a person is seen.

Next, we present the structure of the following related work.

Section 2.4 Handcrafted feature extraction methods: This means *features* based on the human intuition or *learning based features* are used to describe a person by a feature vector. Besides, *metric learning* is introduced to allow for comparison of feature vectors by employing a particular metric that is learned based on the target domain data statistics. Note, in this section, no deep learning based methods are reviewed.

Section 2.5 Data-driven feature extraction methods: *Deep learning* based *feature extraction* is presented. These methods are learned end-to-end from a huge amount of training data. Particular *architectures* are employed to let a network itself decide how to extract features by optimizing a so-called *loss-function*.

Section 2.6 View specific methods: This related work comes closest to our work. Thus, in this section, we discuss related handcrafted and deeply-learned designs where the focus is on exploiting view information. The methods are discussed in more detail than the remaining ones.

Section 2.7 Re-ranking based methods: As post-processing step after a first ranking, a re-ranking can be employed to refine the matching results, and, thus, the person re-identification performance can be improved further.

Section 2.8 Domain adaptation methods: For deploying learned features in open spaces, a domain adaptation method is needed to transfer features learned with one dataset to another one and to improve the generalization characteristics.

Besides, we point out that an overview of person re-identification datasets is given in Appendix A.2.

2.4 Handcrafted feature extraction methods

Early PRID works can be clustered into *feature design* and *metric learning*. Robust *features* are a mandatory prerequisite for successful appearance-based PRID, and metric learning can further boost the re-identification performance by a feature vector comparison that uses a particular learned distance metric instead of employing, e.g., the Euclidean distance or the Cosine distance to compare vectors.

[Gheissari et al., 2006] propose a spatial-temporal segmentation method to detect stable foreground regions of imaged persons. For stable regions, a region-feature-vector consists of i) the

hue and saturation of the region, and ii) the so-called *histogram of edgels* is applied, which is extracted from inside a region. [Gray & Tao, 2008] use color channels of different color spaces (YCbCr, HSV, and RGB [Szeliski, 2011]), 19 texture channels, *Gabor* filters along with *Schmid* filters [Szeliski, 2011], and AdaBoost [Freund & Schapire, 1997] to find the best feature representation from data, instead of designing a feature representation by hand. In conclusion, a performance of 11% rank#1 is found on the VIPeR dataset [Gray et al., 2007] (see Appendix A.2). *Symmetry-Driven Accumulation of Local Features* (SDALF) [Farenzena et al., 2010] segments the foreground of imaged persons from the background, and for each so-called *meaningful part* a symmetry axis is computed. Based on a *meaningful part*, i) the so-called *weighted color histogram* is calculated, which weights pixels near the symmetry axis more than ones which are further away, and forms a color histogram. ii) The so-called *maximally stable color regions operator* detects a set of blob regions to describe features (color, area, and centroid). iii) The so-called *recurrent high-structured patches* are calculated to obtain texture features that are highly recurrent in the appearance of persons. In conclusion, a performance of approximately 19% rank#1 is found on the VIPeR dataset. [Mignon & Jurie, 2012] employ different color spaces (RGB, HSV, YCbCr) for color histograms as well as texture histograms based on the so-called *Local Binary Patterns* (LBP) [Szeliski, 2011] to build a feature vector. Moreover, they propose *Pairwise Constrained Component Analysis* (PCCA) for learning distance metrics from sparse pairwise similarity/dissimilarity constraints in the high dimensional feature space. The PCCA learns a projection into a low dimensional feature space where the distance between the same persons is small, and the distance between different persons is significant. In conclusion, a performance of 19.3% rank#1 is found on the VIPeR dataset. [Köstinger et al., 2012] propose the *Keep It Simple and Straightforward MEtric* approach (KISSME) for *metric learning*. KISSME is based on a log-likelihood ratio test. It is discussed in detail in Section 3.2.1; A performance of 20% rank#1 is found on the VIPeR dataset. [Pedagadi et al., 2013] extract so-called *color histograms* and *moments* (mean, variance, skewness) from images in different color spaces (YUV and HSV) to obtain feature vectors. After that, a dimension reduction follows by using a constrained *principal component analysis* (PCA) to maintain redundancy in color-space. Afterward, the dimensionality is further reduced, using a so-called *Local Fisher Discriminant Analysis* [Szeliski, 2011] defined by a training dataset. A performance of 24.2% rank#1 is found on the VIPeR dataset. [Yang et al., 2014] propose the so-called *salient color names based color descriptor* (SCNCD) and analyze the influence of background and different color spaces. In general, SCNCD can also be stated as applying a color space quantization to the input images and using the quantized colors to describe persons. Yang et al. show that this is more robust against variations in illumination, view, and camera pose. Feature vectors are built, by using a so-called *part based model* to find foreground regions where the SCNCD is calculated. A performance of 37.8% rank#1 is found on the VIPeR dataset. [Liao et al., 2015] introduce the *LOcal Maximal Occurrence descriptor* (LOMO) and include so-called *scale invariant local ternary pattern* (SILTP) histograms [Liao et al., 2010] and color information. Besides, the so-called *Cross-view Quadratic Discriminant Analysis* (XQDA) is presented, which is an extension of KISSME.

Details of XQDA are given in Section 3.2.1. A performance of 40% rank#1 is found on the VIPeR dataset. [Zheng et al., 2015a] employ an 11-dimensional color name descriptor [van de Weijer et al., 2009] for each local patch on an imaged person, and aggregate the descriptors of patches into a global vector exploiting a so-called *Bag-of-Words technique* [Szeliski, 2011]. Furthermore, the large scale dataset Market-1501 is introduced (see Appendix A.2). A performance of 21.7% rank#1 is found on the VIPeR dataset, whereas 42.7% were obtained on Market-1501. [Matsukawa et al., 2016] propose a hierarchical Gaussian (Gaussian of Gaussian "GOG") feature extraction method, which describes the color along with texture cues and models each region by multiple Gaussian distributions. As GOG is state-of-the-art in combination with XQDA, i.e. for handcrafted features in combination with metric learning, it is discussed in detail in Section 3.2.1; a performance of 49.7% rank#1 is found on the small and challenging VIPeR dataset.

In conclusion, the results of conducted experiments indicate that it is useful to train for the target data statistics, and the best method, GOG, in combination with XQDA, can re-identify around the half of the persons from a challenging test dataset.

2.5 Data-driven feature extraction methods

In benefiting from the rapid development of deep learning and from larger PRID datasets, learning-based methods have become increasingly popular since 2014.

Image pairs as training input: [Yi et al., 2014] divide input images into three non-overlapping horizontal parts (approximately consisting of head, body, and legs), and feed the same parts of one probe image and one gallery image through a *Siamese network* [Bromley et al., 1993] during training to obtain a similarity score. The scores from the three parts are fused using a *fully connected layer*, and the network is optimized with back propagation along with a loss-function. During testing, for each input image a feature vector is obtained, which is then compared with other feature vectors from other images by the Cosine distance. A performance of 28.2% rank#1 is found on the VIPeR dataset. [Li et al., 2014] introduce a new large scale dataset CUHK03 (see Appendix A.2) and propose a so-called *filter pairing neural network* (FPNN) to jointly handle misalignment, photometric and geometric transformations, occlusions and background clutter. [Varior et al., 2016] integrate the so-called *long short-term memory* (LSTM) [Hochreiter & Schmidhuber, 1997] into a Siamese network. In the proposed architecture, the LSTM processes image parts sequentially, consequently so-called *spatial connections* can memorize color and texture parts to enhance the discriminative ability of the deep-features. The goal of Varior et al. is to focus more on detailed image information, which, according to the authors, could be overlooked by previous methods. A performance of 37.8% rank#1 is found on the VIPeR dataset, on Market-1501, 62.3% rank#1 is shown.

Image triplets as training input: [Cheng et al., 2016] adapt a so-called *triplet loss function* from the face recognition domain [Schroff et al., 2015], which considers three instead of two images during training. Details of a triplet loss function are given in Section 3.2.1. A performance of 47.8% rank#1 is found on the VIPeR dataset. [Hermans et al., 2017] present the TriNet which employs a novel *within batch hard-negative mining method* and a soft margin for the triplet loss function. Details are also given in Section 3.2.1. The authors show that hard-negative mining is crucial to train such an architecture for person re-identification. On the Market-1501 dataset 84.9% rank#1 is found by applying a hard-negative mining, whereas by using all samples 80.7% rank#1 is obtained.

Explicit handling of image background: [Song et al., 2018] propose a so-called *mask-guided contrastive attention model*. With the help of background masks, features are separately learned from the foreground and background. Besides, a new region-level-triplet-loss is presented, which considers the foreground and background of an image. On the Market-1501 dataset, 83.8% rank#1 is found. [Tian et al., 2018] present a set of experiments to study the background influence on PRID. They propose a so-called *person-region guided pooling deep neural network*, which is based on human parsing maps to learn discriminative *person-part features*. Moreover, a background augmentation technique is presented. A performance of 51.9% rank#1 is found on the VIPeR dataset, whereas on Market-1501 dataset 81.2% rank#1 is shown.

Explicit usage of image sequences (videos): [McLaughlin et al., 2016] propose a Siamese network with *long short-term memory* (LSTM) to match video sequences. First, temporary features are extracted frame-wise by additionally using optical flow, afterwards, recurrent layers and temporal pooling are employed to extract sequence feature vectors. More details are given in the fundamentals, see Section 3.2.2. [Li et al., 2018] elaborate a so-called *spatio-temporal attention model* that automatically discovers a diverse set of distinctive body parts in video sequences. The spatial attention model organizes local image features and combines them by temporal attention. According to the authors, the network alone learns to detect distinctive body parts (body key points such as arm or leg). On the MARS dataset (see Appendix A.2) 82.3% rank#1 and 65.8% mAP are shown, whereas for the "simple" TriNet method, which does not use any body part information nor temporal information, a performance of 79.8% rank#1 and 67.7% mAP are found by considering single images instead of videos on the same dataset. Employing the TriNet including a re-ranking method (we will discuss this later in this Section 2.7) improves rank#1 to 81.2% and 77.4% mAP. Thus, a significant performance difference is shown (+ 1% rank#1 and -11.6% mAP). More training data is needed to evaluate the performance difference between both methods to understand why the contribution of Li et al., only slightly outperforms the TriNet in case of rank#1. [Chen et al., 2018] also deal with image sequences. Instead of using continuous image sequences, they are divided into so-called *snippets* of 8 frames, and the so-called *top-ranked snippets similarities* are aggregated by a network. According to the authors, the intra-person visual variation is minimized, while appearance and temporal information are maintained. Experiments are conducted with and without the usage of optical flow, and different

snipped strategies are evaluated. Finally, a performance of 86.3% rank#1 and 76.1% mAP are found on the MARS dataset.

Person pose based strategies: Another research direction addresses appearance-based PRID in combination with the person pose. Body keypoints are exploited to compare regions of same semantic content such as knee or elbow. Since we do not apply such a technique in our work, we only provide a small overview. [Cheng & Cristani, 2014] use pose information to adapt a so-called *pictorial structure*, i.e. images of persons are partitioned into six parts (chest, head, thighs and legs) to locate the body parts. The so-called *Maximally Stable Color Regions operator* (MSCR) is employed to extract blobs from the body parts. MSCR and color histograms are used to describe an imaged person by a feature vector. [Su et al., 2017] utilize a similar approach, applying an affine transformation to body parts to obtain a unified pose, and the authors present a so-called *Pose-driven Deep Convolutional model* to extract features. Additionally, Su et al. present feature weighting to learn adaptive feature fusions. [Eberle, 2018] present a pose-sensitive embedding. The *view* and fine-grained human poses are used by a CNN to learn a feature representation. This work relies on confidence maps instead of localizing body parts. More details about this method are also given in the next section.

2.6 Person view specific methods

This section deals with methods which are closest to our work. With the term *view specific PRID* we describe the field where view properties are used to obtain view specific feature vectors; or view properties are employed to constrain the matching process.

[Bak et al., 2014] propose a first view specific method. A so-called *pose orientation*, which is basically the horizontal angle around the longitudinal axis of a person compared to the front view, is used in a multi-shot strategy. The method can be summarized as follows: i) imaged persons are oriented upright by an affine transformation before the features are extracted. This allows the reduction of the effect of *perspective-distortion*, i.e., the effect of different camera orientations, e.g., when the camera is tilted. ii) The pose-orientation is determined: Results of person detection and tracking are transformed into a ground-plane coordinate system by using the camera orientation. Then the angle between the moving direction of a person, which is derived from tracking and the camera position in the ground-plane coordinate system, is employed to estimate the *pose orientation*. iii) The *pose orientation* estimations from a video sequence are filtered, and the trajectory of the person is clustered into multiple parts where different person appearances are determined by the filtered pose orientation. iv) From the detected and cropped persons in the images, the background is eliminated and color features are extracted using a weighting strategy that focuses on the foreground along with employing the so-called *Epanechnikov kernel* [Epanechnikov & Seckler, 1969]. v) To measure the similarity between two sets of imaged persons (i.e. between the probe and a person from the gallery), the features

are compared with respect to the *pose orientation*, and the so-called *averaged Epanechnikov weights*. In conclusion, 23% rank#1 performance is obtained, on the SAIVT-Softbio dataset (cf. Appendix A.2).

[Cho & Yoon, 2016] present the so-called *Pose-aware Multi-shot Matching* (PaMM). Here, a *view* is called *pose*. The method can be summarized as follows: calibrated cameras are employed to estimate the *view* of a person for multi-shot matching. i) Assuming that a person mostly walks forward, the view is estimated, similarly to [Bak et al., 2014], as a horizontal angle around the longitudinal axis of a person (continuous value), which is ii) filtered to reduce outliers and occlusions, and iii) the angle is quantized to four view-groups (front, right, back, left). When an image sequence of a person in one camera is given, features are extracted and the view information is estimated. According to the authors, any feature extraction method could be used; exemplary the so-called *dColorSIFT* [Zhao et al., 2013], and the KISSME metric learning method are employed. For the matching of different view combinations, view specific weights are trained using a *Support Vector Machine* [Cortes & Vapnik, 1995] and a training dataset. This means matching of a front–front view or back–back view will receive higher weights than an across-view matching, such as e.g., front–side. To compare two imaged persons, for each view-group (front, right, back, left), the median feature vector is selected for each person. According to the authors, this is reliable, since *"it reflects the characteristics of each group robustly to outliers and furthermore it keeps details"* [Cho & Yoon, 2016, p. 1358]. Afterwards, the distances between the probe and a gallery person are pre-calculated across the median feature vectors of the view groups. Then, to determine a score which indicates the similarity between the two persons, the pre-calculated distances are aggregated with respect to the trained view specific weights, and the resulting values are normalized by the weights. In a consecutive work [Cho & Yoon, 2018] the authors propose a modified matching scheme. Instead of only using the median feature vectors of a view-group, all feature vectors are considered. Unquestionably, the PRID performance of the method depends on the correlation between the training dataset and testing dataset, since the view weighting is learned from data and will likely differ between seasons and use-cases. More importantly [Bak et al., 2014], [Cho & Yoon, 2016], and [Cho & Yoon, 2018] share the fundamental assumption that persons move forward to estimate the view based on the moving direction. However, this is a restriction and does not fit all scenarios. One example is a shopping mall, where persons also tend to move sidewards, along the shelves, thus, the fundamental assumption to classify the view can fail.

[Eberle, 2018, Sarfraz et al., 2018], address the view for PRID (cf. also Sec. 2.5 "Pose based methods"). What we refer to as "view", the authors call "*simple cue of the person's coarse pose (i.e. the captured view with respect to the camera)*". A novel deep learning based architecture is proposed. The architecture is able to classify i) the view, or ii) the person pose, or iii) both pieces of information, and further uses this prior knowledge to extract i) a view specific or ii) a pose-specific or iii) a view-and-pose-specific feature vector. The method is learned in end-to-end fashion, and internally, views (front, back, side) are classified by a so-called *view-*

predictor. Depending on the predicted view, different network weights are activated to extract the output vector. More details about this method concerning the view, are given in Section 5.5.6. Eberle's experiments obtain 88.2% rank#1 on the Market-1501 dataset by using views only, 82.8% rank#1 by pose only, and 87.7% for a fusion of pose and view. Besides, 66.9% mAP is found for using views only, 61.6% mAP for pose only, and 69.0% mAP for a fusion of pose and view. The results indicate that the view performs better than the pose, and for rank#1 the view performs better than a fusion of pose and view. However, for mAP, the fusion of both gives the best performance. Further experiments show that for the Duke dataset (see Appendix A.2), the usage of pose and view gives the best performance for both measures, rank#1 and mAP. Besides, on the Duke dataset the view also performs better than the pose. More importantly, this method does not assume that persons are only moving forward, thus, the method allows the handling of arbitrary moving directions of persons.

[Sun & Zheng, 2019] study the view, here called the *"rotation-angle"*. A large scale synthetic dataset, called *PersonX*, is introduced in which each person of the 1266 persons can be analyzed in horizontal views of 0° to 360° with 10° increments. The effect of view on person re-identification is investigated. Even though a valuable contribution and until today the largest dataset to investigate view information, the synthetic samples seem to be too simple and likely indicate the overfitting of a network, which is underlined by a rank#1 close to a perfect matching. Amongst other methods, Sun & Zheng also use the TriNet for performance comparison. The margin between around 67% rank#1 for real data and close to 95% for synthetic data indicates that the applied model of photo-realism is probably not sufficient. However, according to the authors high performance does not mean the dataset is "easy"; rather, it excludes the influence of environmental factors. Note, this was precisely the goal of their dataset design.

From a general level, all these view specific methods share the same insight, the matching of the same views leads to better performance. Although the methods are inherently designed to handle different views for a better view specific matching, there is no attempt to increase the number of views per individual in the respective camera during system run time to compare views which inherently show the same appearance, i.e. back to back views, side to side views and front to front views.

However, from our point of view, this is mandatory, and it is further underlined by the findings of [Sun & Zheng, 2019], because for approaches that are inherently designed to match arbitrary views, the ambiguities naturally increase if highly asymmetric appearance is presented. This is specifically because rotation variant appearance around the longitudinal axis of a person usually occurs in realistic scenarios.

2.7 Re-Ranking based methods

A further group of methods tackles *re-ranking* as a post-processing step after a first ranking was performed. Re-ranking methods can usually improve the re-identification performance further, if multiple samples per person are provided by each respective camera, e.g., by recording with higher frame rates at the expense of higher computational costs and shorter exposure times. However, in multiple images of the same person, the methods typically assume a similar appearance regardless of the view. We only provide a small re-ranking overview since we do not apply re-ranking in this work.

[Leng et al., 2015] address querying a found gallery match for a given probe person image in a new gallery composed of the original probe and the other gallery images. According to the authors, the same images of persons should not only consist of the same visual content, but rather also possess similar k-nearest neighbors in the ranking list. [García et al., 2015] propose a ranking optimization method based on discriminant context information analysis. The method refines an initial ranking by removing the visual ambiguities common to first ranks. To do so, content and context information is analyzed. [Sarfraz et al., 2018] present *Expanded Cross Neighbourhood Re-Ranking*. The method is unsupervised and does not need to compute a new rank list for each image pair.

2.8 Domain adaptation methods

The high performance of recent methods stems from training in the target domain. Unsupervised learning strategies or domain adaptation strategies inherently show a low re-identification performance compared to the performance obtained for supervised training in the target domain. However, for practical application, it can be a tremendous task to collect a new dataset of each target domain. Since we do not address domain adaptation techniques in this thesis, merely a small overview is given: [Xiao et al., 2016] propose so-called *Domain guided Dropout*, a technique where different datasets are combined to train a model by using a dropout technique for neurons. Here, neurons which are merely important for particular datasets are dropped, and, thus, the drop of unneeded neurons helps the network to generalize better. [Bak et al., 2018] argue that illumination changes across cameras makes PRID challenging, and, thus, propose an unsupervised domain adaptation technique by using synthetic images and a Generative Adversarial Network (GAN) [Goodfellow et al., 2014] to bridge the domain gap for PRID.

2.9 Discussion

General discussion: The appearance-based approaches typically share the issues that they inherently suffer from different scene illumination, occlusions, changes in the camera pose, background clutter, and image quality. Recall Figure 1.2, where a moderately difficult sample was illustrated. If one person view in the gallery is the opposite person view as that in the probe, it is finally a question of dataset size and the diversity level of persons if a correct match can be found. Therefore, in this work, we elaborate an approach, which i) is for the first time able to provide different views for each person in the respective camera, and ii) this view information is used as strong prior knowledge in a subsequent fusion method.

Discussion about the related work: Today, end-to-end learning methods allow for achieving of a rank#1 performance up to 95% on the Market-1501 dataset, whereas handcrafted methods obtain around 50% rank#1.[5] Most works apply supervised learning on the respective training dataset and report the performance obtained by using the corresponding test dataset. Applying this method does not allow the conclusion that the network can generalize or is only overfitted with respect to the particular test dataset. Thus, high performances should not be interpreted as the re-id task being solved. Rather, the networks can solve a particular test dataset well, which has similar data statistics to the training dataset. By considering that the complexity of available datasets only presents a small excerpt of the real world (cf. Appendix A.2), the generalization ability is not finally answered, and, to the best of our knowledge, no product is available which successfully and reliably provides appearance-based re-identification. Many open research questions are unanswered, such as: i) Certainty: which detail level between two different persons in image space can be distinguished. ii) Generalization ability: how to employ domain transfer without the need to create a fully annotated dataset for end-to-end learning for each new scenario. iii) Application transfer: the majority of related work focuses on the closed-set. However, how can we transfer this knowledge about *similarity ranking* between imaged persons to tackle the open-set task?

[5] Leader board, see: `http://www.liangzheng.com.cn/Project/state_of_the_art_market1501.html`

Chapter 3

Fundamentals

In this chapter, the fundamentals, which are used in our approach, are introduced. These are fisheye cameras and feature extraction methods for imaged persons.

3.1 Fisheye camera geometry and projection model

The projection of a scene into an image can mathematically be approximated by a projection model, which describes the projection process by the usage of light rays. Several projection models exist. For central projection, the simplest model is the pinhole camera model, which can be extended by considering distortions [Brown, 1966]. For special-purpose projections, such as for omnidirectional cameras, other projection models are mandatory to consider the geometry of the optical path.

An *omnidirectional camera*, in particular, a fisheye camera, is not popular in the measurement sciences. However, in fields such as robotics or the automotive domain, omnidirectional cameras are typically used to cover a larger "ultra-wide-angle" field of view.[1] These cameras can be built as dioptric cameras, where a combination of shaped lenses are used; a catadioptric camera, where a combination of a standard camera and mirror is applied; or a polydioptric camera, where multiple cameras with overlapping fields of view are leveraged [Scaramuzza, 2014].

A fisheye camera, which consists of an ordinary camera in combination with particular lenses, is a dioptric camera. Here, lenses with a large negative meniscus element, which are mounted on the head of a compact positive component, are used to create a lens-system with a long back focal distance[2] and a short focal length [Thibault, 2010]. These lens-systems are favorable for a wide-angle field of view [Kingslake, 1985]. An easy to note property of a fisheye projection is that straight lines in object space are not straight in fisheye image space, but rather parts of

[1] Field of view: wide-angle $\measuredangle \geq 60°$ [Thibault, 2010]; ultra-wide-angle $\measuredangle \geq 180°$ [Scaramuzza, 2014]

[2] This is the mechanical distance between the mating surface of a lens on the housing and the sensor.

an ellipse. For such ultra-wide-angle lenses that cover a hemispherical field the pinhole camera model is not valid anymore.

By using central projection the distance r_{CP}, between the principal point and a point in space projected onto an image plane, can be approximated to $r_{CP} = f \cdot tan(\theta)$ [Förstner & Wrobel, 2016], where f is the focal length and θ is the angle between the point in space and the optical axis as seen from the camera origin. However, the distance approaches infinity for θ approaching 90°, which needs, in theory, huge image sensors to project a scene on the image plane.

On the contrary, using e.g. the equidistant fisheye projection, the corresponding distance, of the same point in space, is given by $r_{FE} = f \cdot \theta$ [Förstner & Wrobel, 2016]. Due to the linear relationship between focal length and the angle, the projection enables ultra-wide-angle projection on a plane. Moreover, it is important to note that the inherent large distortions of a fisheye lens are not the result of aberration. Instead, they are the result of projecting a hemispheric field on a circle on the image plane, which is not possible without distortions [Thibault, 2010].

Following [Schönbein, 2014], many projection models exist to map a point from the object space into the fisheye image space. Lens vendors select a projection model which they approximate in the manufacturing process, e.g. the mentioned equidistant projection model. [Kang, 2000, Svoboda & Pajdla, 2002] use projection models for particular sensor types. Other fisheye projection models such as stereographic projection, orthogonal projection, equisolidangle projection are reviewed in e.g. [Abraham & Förstner, 2005], [Förstner & Wrobel, 2016]. [Geyer & Daniilidis, 2000] propose the so-called *sphere camera model* which was extended by [Barreto & Araújo, 2001]. The *sphere camera model* allows for efficient forward and backward projection and unifies all central catadioptric cameras in a two-step projection. [Ying & Hu, 2004] show that this model is also valid for fisheye projection. The *Mei Model* [Mei & Rives, 2007] extends the model with a perspective lens and additionally models distortions for misalignment between the mirror and the camera axis.

Following the notations of Mei and Rives (cf. Fig. 3.1), points from object space ($\mathbf{X} = [X, Y, Z]^T$) are projected into the FE image plane ($\mathbf{x}_{FE} \in \pi_{FE}$) as follows:

1) Using the centre of a unit sphere as coordinate origin (cf. coordinate system $\mathcal{F}_m$ in Fig. 3.1), points are projected onto that sphere,

$$(\mathbf{X})_{\mathcal{F}_m} \rightarrow (\mathbf{X}_S)_{\mathcal{F}_m} = \frac{\mathbf{X}}{||\mathbf{X}||} = (X_S, Y_S, Z_S). \tag{3.1}$$

2) A new coordinate system, $\mathcal{F}_p$, is introduced, which compared to $\mathcal{F}_m$, is translated by ξ to yield $\mathbf{C}_p = (0, 0, \xi)$, where the size of ξ depends on the employed lens,

$$(\mathbf{X_S})_{\mathcal{F}_m} \rightarrow (\mathbf{X_S})_{\mathcal{F}_p} = (X_S, Y_S, Z_S + \xi). \tag{3.2}$$

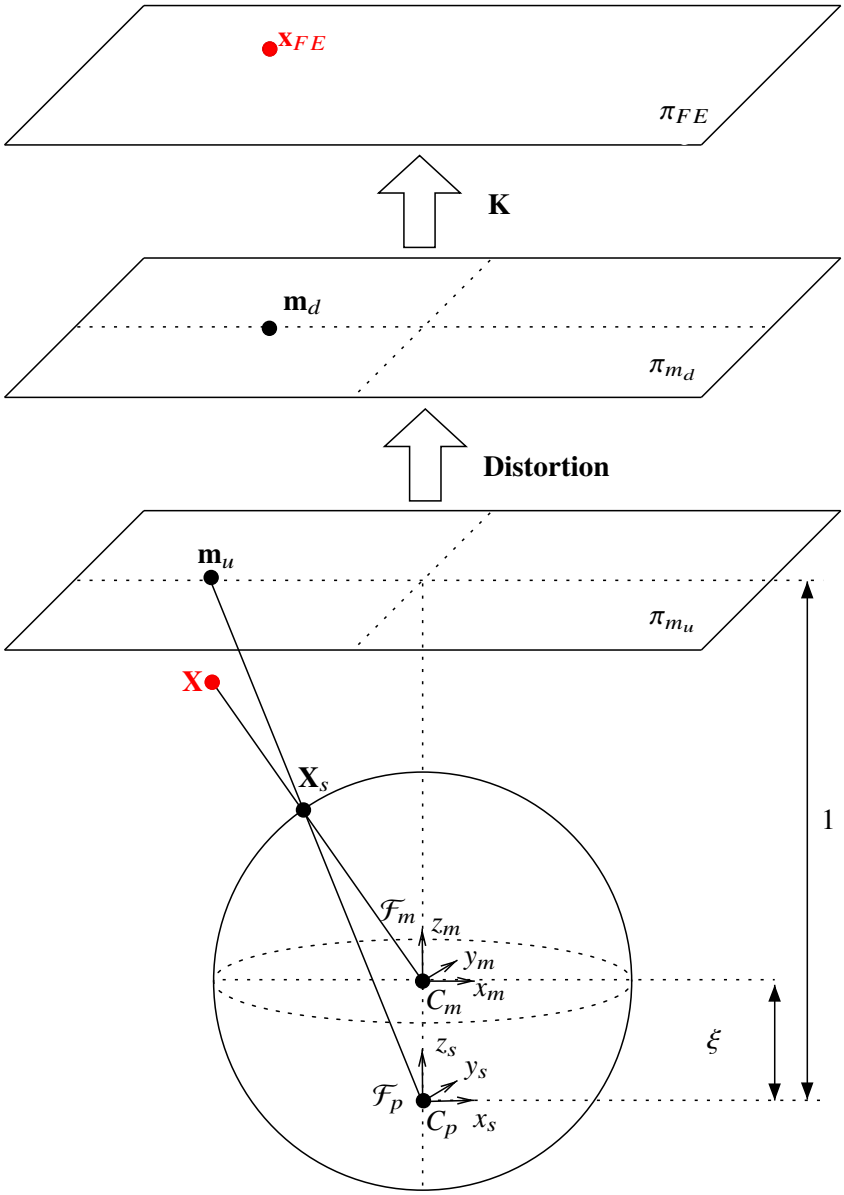


Figure 3.1: Projection model from [Mei & Rives, 2007] adapted for this thesis. Note that the radius of the unit sphere and the distance between the unit sphere center and the plane π_{m_u} are both "1", since the plane is the tanget plane of the sphere. In the illustration, for reasons of simplicity, the plane π_{m_u} is shifted.

3) The points are projected onto the image plane (π_{m_u}), which is the tangent plane of the sphere. Thus, the normalized coordinates on the image plane are given by

$$\mathbf{m}_u = \left(\frac{X_S}{Z_S + \xi}, \frac{Y_S}{Z_S + \xi}, 1\right). \tag{3.3}$$

4) According to [Mei & Rives, 2007] radial and tangential distortions [Brown, 1966] are added. The coordinates on the distortion affected image plane (π_{m_d}) read:

$$\mathbf{m}_d = \mathbf{m}_u + D(\mathbf{m}_u, \mathbf{V}), \tag{3.4}$$

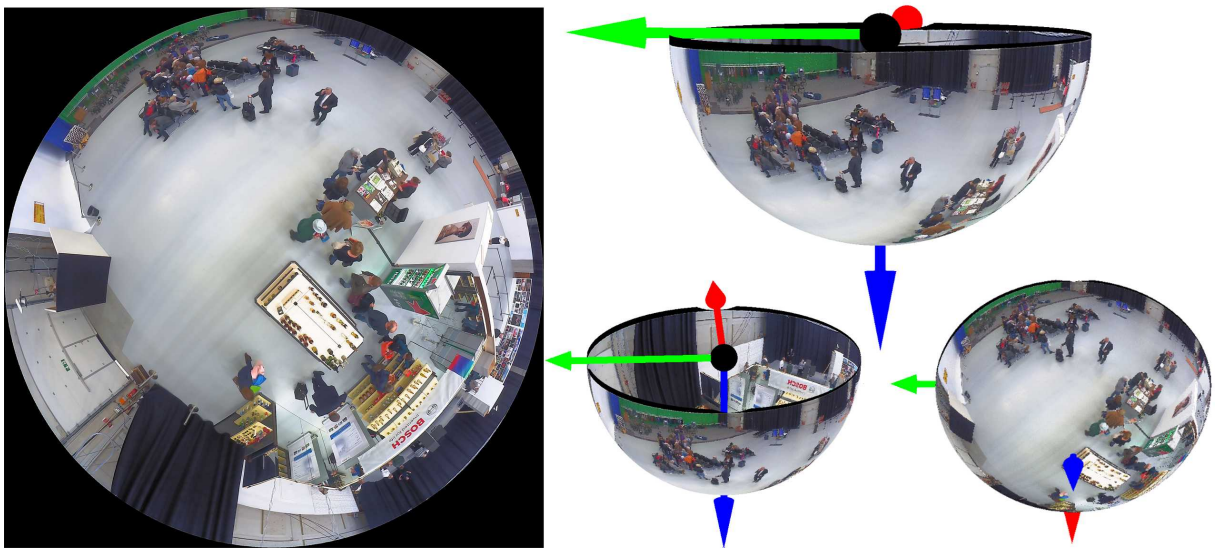


Figure 3.2: Fisheye image and de-warped illustrations. Left: visualization of one recorded circular fisheye image where the camera is mounted in nadir orientation. Right: three different illustrations where the fisheye image is back-projected onto the half-unit-sphere (cf. Fig. 3.1). The three examples are recorded by three virtual cameras of different orientations. For better understanding, the colored arrows indicate the same fixed coordinate system. Figure inspired by [Knorr, 2018].

where D describes the coordinate dependent distortion with the distortion coefficients $\mathbf{V}$. Note that the fisheye distortion is by itself already a function of the radial displacement, so the fisheye projection model will inevitably "absorb" a major part of the distortion. This means, depending on the required level of accuracy and use case along with manufacturing precision, D can sometimes be neglected.

5) The final projection involves a generalized camera projection matrix

$$\mathbf{K} = \begin{pmatrix} c & c \cdot s & x_0 \\ 0 & c \cdot r & y_0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.5}$$

with c the generalized focal length, $[x_0, y_0]$ the principal point coordinates, s the skew, and r the aspect ratio. The coordinates in the fisheye image plane are finally given by

$$\mathbf{x}_{FE} = \mathbf{K} \cdot \mathbf{m}_d. \tag{3.6}$$

Section 4.3 describes, how fisheye image cut-outs are transformed to central projection geometry of a virtual camera. According to the derived coordinates, the pixel values of a virtual image in a particular intermediate coordinate system (see steps 1-5) can be obtained via an indirect coordinate-to-pixel-intensity look-up table. For a general view, Figure 3.2 shows one fisheye image and the same back-projected onto the unit sphere, which is observed from three virtual cameras of different orientations.

3.2 Feature extraction

In our approach, we integrate a feature extraction method from related work. This feature extraction method is one of several modules in our pipeline, and can be replaced by almost any other method; different strategies are evaluated in the experimental part of this thesis (cf. Chapter 5). In order to understand the basic idea and process, one handcrafted method, and two data-driven methods (based on deep learning), which are used later, are described in this section. All show high state-of-the-art performance for the respective feature design (handcrafted or data-driven).

3.2.1 GOG/XQDA - a handcrafted feature extraction method

To derive handcrafted features, prior knowledge is employed in the feature design. In addition to selecting good handcrafted features, the PRID performance can further be improved by the use of metric learning. GOG, one state-of-the-art handcrafted feature extraction approach, that we apply in in combination with the metric learning method XQDA, and a detailed description of metric learning are given in this section.

GOG: The selected handcrafted feature extraction method is *Hierarchical Gaussian Descriptor for Person Re-Identification* [Matsukawa et al., 2016, Matsukawa et al., 2019] which for handcrafted features shows state-of-the-art performance for a broad range of datasets, and in particular among all methods superior performance for small datasets such as VIPeR [Gray et al., 2007]. In a first step (cf. Fig 3.3), GOG divides person images of equal size into $G = 7$ overlapping horizontal stripes, called regions.

In general, a region is modelled as a set of multiple Gaussian distributions. In each of them, the appearance of a local patch is described. The characteristics of a set are again described by another Gaussian distribution. In contrast to previous methods GOG jointly includes mean and covariance, which according to Matsukawa et al. is beneficial in describing texture and color information simultaneously and remedies the effect of noise and spatial alignment.

Consequently, in the second step, after an image is divided into regions, for each region, squared patches (5×5 pixels) are densely extracted with $p = 2$ pixel interval. To describe each pixel i of a patch, an intermediate feature vector

$$\mathbf{\Phi}_i = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, R, G, B]^T \tag{3.7}$$

is extracted, where y is the pixel location in the vertical direction, $M_{\Theta \in \{0^\circ,\ldots,270^\circ\}}$ are the magnitudes of the pixel intensity gradient along the orientations, and "R,G,B" are the color channel values. Each dimension of $\mathbf{\Phi}_i$ is linearly mapped to the range $[0, 1]$ to normalize the scales of different feature values. The pixel locations are used to exploit spatial information within each

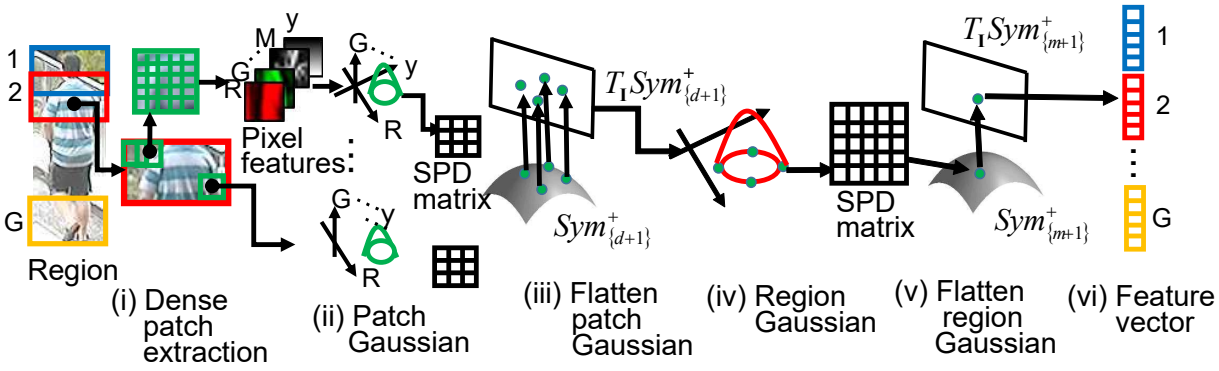


Figure 3.3: GOG descriptor: (i) For each region, local patches are densely extracted. (ii) each of the local patches is described via a Gaussian distribution of pixel features; these Gaussians are referred to as patch Gaussians. (iii) Each patch Gaussian is flattened and vectorized by considering the underlining geometry of Gaussians. (iv) Then, the patch Gaussians inside a region are aggregated into a region Gaussian. (v) The region Gaussian is further flattened and the feature vector is created. (vi) Finally, from all regions, the feature vectors are extracted and concatenated into one vector. © Reprinted, with permission, from [Matsukawa et al., 2019].

region, whereas gradient information is exploited to describe textural information, and color information gives an essential cue for PRID. After feature extraction inside a patch s the features are modeled using a Gaussian distribution

$$\mathcal{N}(\Phi; \mu_s, \Sigma_s) = \frac{1}{(2\pi)^{d/2}|\Sigma_s|} e^{(-\frac{1}{2}(\Phi-\mu_s)^T \Sigma_s^{-1}(\Phi-\mu_s))}, \tag{3.8}$$

where $|\cdot|$ is the determinant of a matrix, d is the feature vector dimension, μ_s is the mean vector and Σ_s is the covariance matrix of the sampled patch s. This Gaussian from inside a patch is called patch Gaussian.

Next, as "Gaussian of Gaussian" indicates, the sets of patch Gaussian, from inside a region, are modelled by another Gaussian. For the aggregation, mathematical operations such as mean and covariance of the Gaussian are required which need a mathematical discussion:

"From the viewpoint of information geometry, the space of probability distribution is considered as a Riemannian manifold where the Euclidean operation cannot be applied directly [Amari & Nagaoka, 2000]. A Riemannian manifold can be locally flatten(ed) into a Euclidean space by projecting it into a tangent space endowed with Riemannian metric. The space of the Symmetric Positive Definite (SPD) matrix is also considered as a Riemannian manifold and this space is recently well understood. The log Euclidean metric [Arsigny et al., 2006] for SPD matrix provides a solid way to map a point on the manifold to a Euclidean tangent space via a matrix logarithm." [Matsukawa et al., 2016, p. 1366]

Therefore, Matsukawa et al. propose applying a tangent space mapping and a half vectorization. To exploit the benefit of the log Euclidean metric [Arsigny et al., 2006], the patch Gaussian is

embedded in the Symmetric Positive Definite (SPD) matrix manifold (Sym^{+}_{d+1}) [Lovric et al., 2000] of dimension $d + 1$. The patch Gaussian $\mathcal{N}(\mu_s, \Sigma_s)$ is projected into Sym^{+}_{d+1} as $\mathbf{P_s}$ by:

$$\mathcal{N}(\Phi; \mu_s, \Sigma_s) \sim \mathbf{P}_s = |\Sigma_s|^{\frac{1}{d+1}} \begin{bmatrix} \Sigma_s + \mu_s \mu_s^T & \mu_s \\ \mu_s^T & 1 \end{bmatrix}. \tag{3.9}$$

Afterwards, the log Euclidean and half vectorization are applied [Arsigny et al., 2006]

$$g_s = vec(log(\mathbf{P}_s)), \tag{3.10}$$

where $log(\bullet)$ is the matrix logarithm operator, and $vec(\bullet)$ takes the upper triangular part of the symmetric matrix as vector.

In different images, the positions of local parts vary, due to person and camera pose variations. The presented solution aggregates the local patches into an orderless representation. In particular the flattened patch Gaussians are aggregated in a region distribution, which is modeled by a Gaussian distribution. To partly suppress the effect of outer image regions which are assumed to be background, weighting is introduced for each patch

$$w_s = exp(-(x_s - x_c)^2/2\sigma^2), \tag{3.11}$$

where $x_c = W/2$, $\sigma = W/4$, W is the image width, and x_s is the x-coordinate of the center pixel of patch s. The weighted mean vector and covariance matrix read:

$$\mu^{\mathcal{G}} = \frac{1}{\sum_{s \in \mathcal{G}} w_s} \sum_{s \in \mathcal{G}} w_s g_s \quad , \quad \sigma^{\mathcal{G}} = \frac{1}{\sum_{s \in \mathcal{G}} w_s} \sum_{s \in \mathcal{G}} w_s (g_s - u^{\mathcal{G}})(g_s - u^{\mathcal{G}})^T, \tag{3.12}$$

where $\mathcal{G}$ is the region in which the patch Gaussians are aggregated. The region is modeled as $\mathcal{N}(g; \mu^G, \Sigma^G)$.

Afterwards the regions are flattened, to match among region descriptors in Euclidean space. This is applied in the same manner as by Equation 3.9 to $\mathcal{N}(g; \mu^G; \Sigma^G) \sim Q$ where Q is a $(m + 1) \times (m + 1)$ SPD matrix. By extracting the region Gaussian for each of G regions, the feature vectors $\{\mathbf{\Phi}_g\}_{g=1}^{G}$ are obtained. The feature vectors are concatenated, to preserve spatial locations, into the final representation $\mathbf{\Phi}_{RGB} = [\mathbf{\Phi}_1^T, ..., \mathbf{\Phi}_G^T]^T$ which describes the whole person image as one feature vector.

To improve GOG further, Matsukawa et al. propose extending the feature extraction from RGB color space by Lab, HSV, and nRnG colorspaces to a fused feature vector to become more robust due to complementary color space properties. Thus, for each colorspace, the calculations from above are applied, and the resulting feature vectors are concatenated to

$$\mathbf{\Phi} = [\mathbf{\Phi}_{RGB}^T, \mathbf{\Phi}_{Lab}^T, \mathbf{\Phi}_{HSV}^T, \mathbf{\Phi}_{nRnG}^T]^T. \tag{3.13}$$

XQDA: Whereas GOG results in a handcrafted feature vector, a high performance of this method can only be obtained in combination with metric learning. The authors of GOG show that the so-called *Quadatric Discriminant Analysis (XQDA)* [Liao et al., 2015] gives the best performance. Following Liao et al., XQDA can be introduced as follows:

Consider a sample difference $\Delta = \mathbf{\Phi}_i - \mathbf{\Phi}_j$, where Δ is the intra-person difference if $i = j$, and $i \neq j$ is called extra-person difference [Moghaddam et al., 2000]. The two classes of variation can be defined as Ω_i for intra-person variations and Ω_j for extra-person variations.

[Köstinger et al., 2012] apply the log-likelihood ratio test [Neyman & Pearson, 1928] of the two Gaussian distributions to person re-identification. Under zero-mean Gaussian distribution, the likelihoods of observing Δ in Ω_I (Intra) and Ω_E (Extra) can be defined as:

$$p(\Delta|\Omega_I) = \frac{1}{(2\pi)^{d/2}|\Sigma_I|^{1/2}} e^{-\frac{1}{2}\Delta^T\Sigma_I^{-1}\Delta} \quad , \quad p(\Delta|\Omega_E) = \frac{1}{(2\pi)^{d/2}|\Sigma_E|^{1/2}} e^{-\frac{1}{2}\Delta^T\Sigma_E^{-1}\Delta}, \tag{3.14}$$

whereas Σ_I and Σ_E are the covariance matrices of Ω_I and Ω_E, respectively. By employing the Bayesian rule and the log-likehood ratio test, the hypothesis that a pair $(\mathbf{\Phi}_i, \mathbf{\Phi}_j)$ is from different persons is tested versus the alternative:

$$\delta(\Omega) = log\frac{p(\mathbf{\Phi}_i, \mathbf{\Phi}_j|\Omega_E)}{p(\mathbf{\Phi}_i, \mathbf{\Phi}_j|\Omega_I)} = log\frac{p(\Delta|\Omega_E)}{p(\Delta|\Omega_I)}. \tag{3.15}$$

The comparison is made in the space of pairwise differences with zero mean, where a high value of $\delta(\Omega)$ indicates that the hypothesis is validated, otherwise it is rejected. In other words, a high value can be used to obtain a large distance between two feature vectors of different persons. These underlying thoughts can be simplified along with equation 3.15 by logarithmic operations and deleting constant terms, as they merely provide an offset, to:

$$f(\Delta) = \Delta^T(\Sigma_I^{-1} - \Sigma_E^{-1})\Delta, \tag{3.16}$$

and the corresponding distance function between $\mathbf{\Phi}_i$ and $\mathbf{\Phi}_j$ follows as

$$d(\mathbf{\Phi}_i, \mathbf{\Phi}_j) = (\mathbf{\Phi}_i - \mathbf{\Phi}_j)^T(\Sigma_I^{-1} - \Sigma_E^{-1})(\mathbf{\Phi}_i - \mathbf{\Phi}_j). \tag{3.17}$$

This shows that learning the distance function corresponds to estimating the covariance matrices Σ_I^{-1} and Σ_E^{-1}.

Liao et al. extend the method of Köstringer et al. by cross-camera-view[3] metric learning. In particular, a subspace $W = (w_1, w_2, ..., w_n) \in \mathbb{R}^{d\times r}$ is learned with cross-camera-view data,

[3] Note, in [Liao et al., 2015] cross-camera-view is called cross-view since there is no relation to views of a persons. Since we focus on different views in this work, we replaced views with camera-views.

where d is the dimension of the original feature vector, and $\mathbb{R}^r (r < d)$ is the learned subspace. Consider a cross-camera-view training set $\mathbf{X}, \mathbf{Y}$ of c classes where

$$\mathbf{\Phi}_X = (\mathbf{\Phi}_{x_1}, \mathbf{\Phi}_{x_2}, ..., \mathbf{\Phi}_{x_n}) \in \mathbb{R}^{d \times n} \tag{3.18}$$

contains n samples in a d-dimensional space from one camera view.

$$\mathbf{\Phi}_Z = (\mathbf{\Phi}_{z_1}, \mathbf{\Phi}_{z_2}, ..., \mathbf{\Phi}_{z_m}) \in \mathbb{R}^{d \times m} \tag{3.19}$$

contains m samples in the same d-dimensional space from another camera view. Considering a subspace W, the distance function (3.17) in the r-dimensional subspace is computed as

$$d_W(\mathbf{\Phi}_x, \mathbf{\Phi}_z) = (\mathbf{\Phi}_x - \mathbf{\Phi}_z)^T W(\Sigma_I^{'-1} - \Sigma_E^{'-1}) W^T (\mathbf{\Phi}_x - \mathbf{\Phi}_z), \tag{3.20}$$

where $\Sigma_I^{'-1} = W^T \Sigma_I W$ and $\Sigma_E^{'-1} = W^T \Sigma_E W$.

Liao et al. propose to learn the kernel matrix $M(W) = W(\Sigma_I^{'-1} - \Sigma_E^{-1})W^T$, but claim that directly optimizing d_W is difficult. Note, Ω_I and Ω_E have zero mean, then, given a basis w, the projected samples of the classes will still have zero mean. Thus, the variances σ_E and σ_I can be used to distinguish between the two classes. Therefore, the projection direction w can be optimized such that $\sigma_E(w)/\sigma_I(w)$ is maximized. By considering $\sigma_\mu(w) = w^T \Sigma_\mu w$, the *Generalized Rayleigh Quotient* follows as

$$J(w) = \frac{w^T \Sigma_E w}{w^T \Sigma_I w}. \tag{3.21}$$

Then, the maximization of $J(w)$ is used,

$$max_w w^T \Sigma_E w, \text{ such that } w^T \Sigma_I w = 1, \tag{3.22}$$

which can be solved by a generalized eigenvalue decomposition problem, to obtain the final solution ($M(w)$).

In summary, by applying metric learning the distance between two images, one probe $\mathbf{I}_\mathcal{P}$ and one gallery sample $\mathbf{I}_{\mathcal{G}_\mu}$, is first determined by feature extraction method, here GOG,

$$\mathbf{\Phi}_\mathcal{P} = f(\mathbf{I}_\mathcal{P}), \mathbf{\Phi}_{\mathcal{G}_\mu} = f(\mathbf{I}_{\mathcal{G}_\mu}). \tag{3.23}$$

Afterwards, the distance between two images is determined in feature space by

$$D(\mathbf{\Phi}_\mathcal{P}, \mathbf{\Phi}_{\mathcal{G}_\mu}) = (\mathbf{\Phi}_\mathcal{P} - \mathbf{\Phi}_{\mathcal{G}_\mu}) \cdot \mathbf{M} \cdot (\mathbf{\Phi}_\mathcal{P} - \mathbf{\Phi}_{\mathcal{G}_\mu})^T. \tag{3.24}$$

The task of metric learning is to determine the matrix $\mathbf{M}$. The matrix is learned from training data. Note, for the special case $\mathbf{M}$ is an identity matrix the squared Euclidean distance between $\mathbf{\Phi}_{\mathcal{P}}$ and $\mathbf{\Phi}_{\mathcal{G}_\mu}$ is obtained.

Finally, some practical results are discussed for GOG and XQDA. Applying GOG with the RGB color space gives 19.5% rank#1 on the VIPeR dataset [Gray et al., 2007]. Using all color spaces gives 21.1% rank#1. Enabling XQDA allows 49.7% rank#1. For the CUHK01 dataset [Li et al., 2012], rank#1 improves from 19.6%, to 21.7%, to 57.8%. This large performance boost indicates that using data for learning the target data statistic is helpful to obtain a high person re-identification performance. Note, the reference to the rank#1 performance values is the supplementary material of [Matsukawa et al., 2016].

3.2.2 TriNet and SRNN - two data-driven feature extraction methods

Deep learning based methods are becoming more and more important in several domains.[4] An extensive review of related methods for person re-identification was given in Chapter 2.

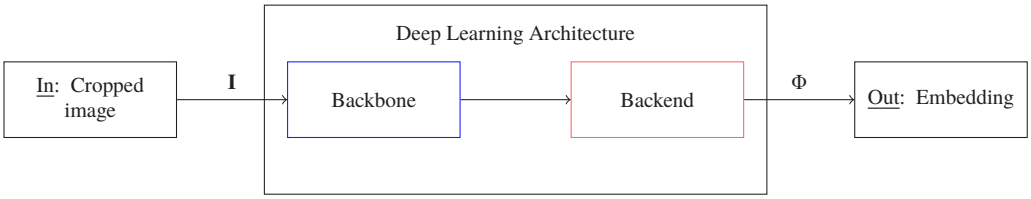
Convolutional Neural Network (CNN) based methods typically have in common (cf. Fig. 3.4a), that an input image ($\mathbf{I}$) is fed through a deep learning based network architecture to obtain an injective representation ($\mathbf{\Phi}$) of the image in latent space. This can be expressed as $f : \mathbf{I} \hookrightarrow \mathbf{\Phi}$.

A common network architecture for person re-identification can be divided into a *backbone* and a subsequent following *backend* (cf. Fig. 3.4a). In the backbone so-called *feature maps* are extracted, whereas the final layer of the backend is used to obtain a so-called *embedding*. Such an embedding, which basically is a feature vector, can be used to compare the similarity between imaged persons by comparing the distances between the vectors. Thus, an embedding is the non-linear processing result of the high-dimensional feature maps in a low-dimensional latent space, which, in addition to the feature map extraction, is learned from data.

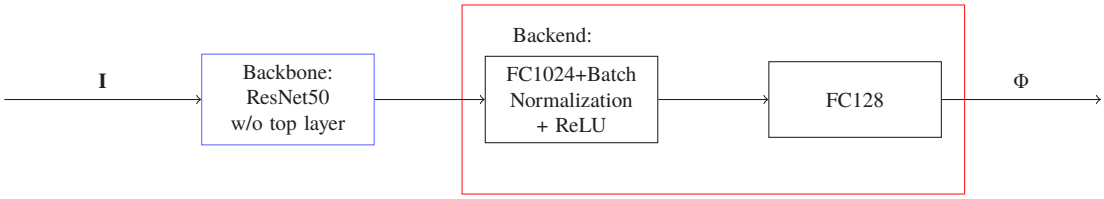
Today, state-of-the-art networks such as ResNet50 [He et al., 2016], VGG16 [Simonyan & Zisserman, 2014], or InceptionV3 [Szegedy et al., 2016] are applied as a backbone. These networks are pre-trained for a particular auxiliary task with a large image dataset, for example, training for image classification using the large scale ImageNet dataset [Deng et al., 2009]. After pre-training, the last layer(s) are removed, and some target-task specific layers are appended, which are typically called backend. Many PRID publications focus on the development of such backends to improve the re-identification, since high-level and re-id specific tasks, such as person pose estimation or alignment, can be jointly modeled here.

TriNet: In the following TriNet, the feature extraction method by [Hermans et al., 2017] is described, which uses a *triplet loss function* and *within batch hard-negative mining* during

[4] For a more detailed theory of deep learning, the reader is referred to the standard literature such as [Bishop, 2007] and [Goodfellow et al., 2016].



(a) General processing of a deep learning based feature extraction method



(b) Processing with TriNet. The backbone uses a ResNet50 without top layer. Two fully connected layers (FC), Batch Normalization [Ioffe & Szegedy, 2015] and rectified linear unit (ReLU) activation [Glorot et al., 2011] are used as backend.

Figure 3.4: Schematic illustration of a deep learning based feature extraction method. **(a)** General processing, **(b)** architecture for the TriNet [Hermans et al., 2017].

training. As depicted in Figure 3.4b a state-of-the-art backbone (ResNet50 [He et al., 2016] without top layer) is applied and two fully connected layers (FC) with 1024 and 128 neurons are used as backend. Thus the embedding is a 128-dimensional feature vector. Between the FC layers batch normalization [Ioffe & Szegedy, 2015] is employed for regularization, and a rectified linear unit (ReLU) [Glorot et al., 2011] is used as activation function.

In the past, triplet loss functions [Weinberger & Saul, 2009] were investigated for different applications, e.g. for face recognition [Schroff et al., 2015]. The geometric interpretation of a triplet loss function can be stated as follows (cf. Fig. 3.5): Given a person image (anchor), the distances between images of one and the same person (positive samples) are minimized during training, typically by back-propagation, while the distances between images of different persons (negative samples) are maximized.

The main contribution of TriNet is a new within-batch hard-negative mining method for the triplet loss function, called *Batch Hard Loss*, and a soft margin version. During learning, P persons and K images per person are randomly selected within a batch.[5] For each person, the hardest positive and the hardest negative sample are further calculated to obtain the loss value $\mathcal{L}_{\text{BH}}$,

[5] Note that the batch size, which is typically dictated by the available hardware, can be calculated by $B = P \cdot K$.

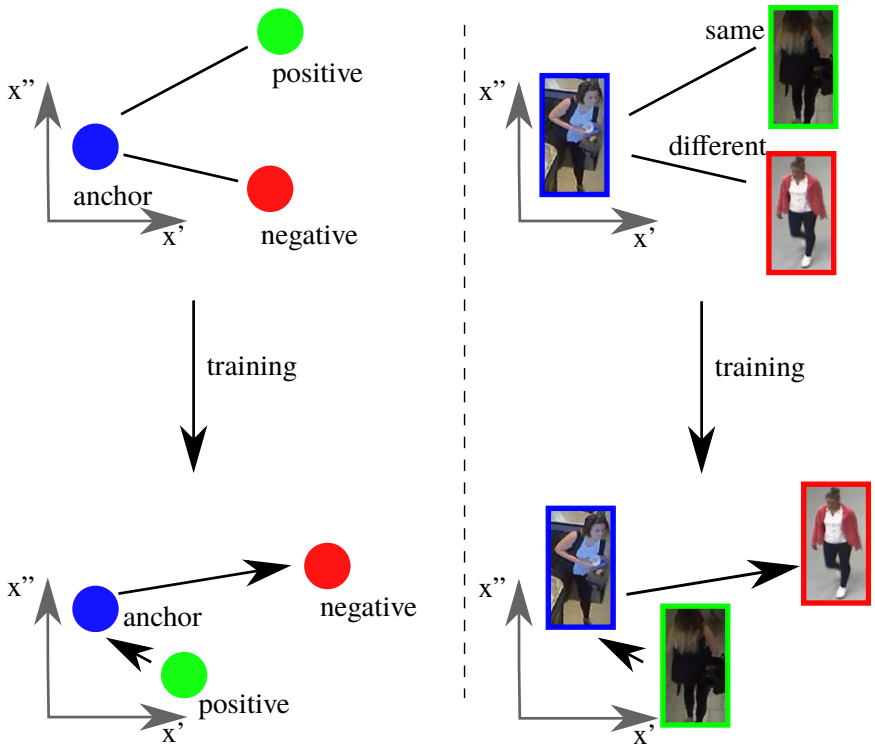


Figure 3.5: Geometric interpretation behind the triplet loss function. The left side shows a simplified example in a two-dimensional plane projection (x’,x”) of the latent space. The right side shows the same with person images. The objective of using the triplet loss is to minimize the distance between an anchor point (blue circle) during training and the same person (positive sample, green circle) and to maximize the distance between the anchor and a different person (negative sample, red circle). (Figure inspired by [Schroff et al., 2015])

$$\mathcal{L}_{\text{BH}} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}} \Big[m + \overbrace{\max_{p=1\ldots K} D\left(\boldsymbol{\Phi}(\mathbf{I}_a^i), \boldsymbol{\Phi}(\mathbf{I}_p^i)\right)}^{\text{hardest positive}} - \overbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(\boldsymbol{\Phi}(\mathbf{I}_a^i), \boldsymbol{\Phi}(\mathbf{I}_n^j)\right)}^{\text{hardest negative}} \Big]_+, \tag{3.25}$$

whereas, $\mathbf{I}_a^i$ is an anchor sample, $\mathbf{I}_p^i$ is a positive sample, $\mathbf{I}_n^i$ is a negative sample, $\boldsymbol{\Phi}(\bullet)$ is the image representation in latent space, and m is a margin value.

Given an anchor sample, a positive sample which belongs to the same person is closer than that of a negative sample by at least the margin m. Thus, the basic idea behind the inner term $[m + \bullet]_+$ is "*to avoid correcting already correct triplets*" [Hermans et al., 2017, p. 4], to pull closer samples from the same class as much as possible [Zhang et al., 2016] [Cheng et al., 2016], and enforce the margin m between positive and negative samples.

In TriNet, $[m + \bullet]_+$ is replaced by a smooth approximation using the *softplus function*: $\ln(1 + \exp(\bullet))$. According to [Hermans et al., 2017], the softplus function behaves in a similar way to the inner term, but it decays exponentially instead of having a hard cut-off.

During inference, the distance D between two person patches, the probe and one gallery sample, is found by using the Euclidean distance in the learned latent space

$$D\left(\mathbf{\Phi}(\mathbf{I}_{\mathcal{P}}), \mathbf{\Phi}(\mathbf{I}_{\mathcal{G}_i})\right) = \left\|\mathbf{\Phi}(\mathbf{I}_{\mathcal{P}}) - \mathbf{\Phi}(\mathbf{I}_{\mathcal{G}_i})\right\|_2 . \tag{3.26}$$

As depicted in the right part of Figure 3.5, two aspects arise from using such a feature extraction method. i) Domain gap: the network learns, based on a training dataset, how to map an image into the latent space, which can lead to a bias for general use cases if the correlation between training and test data is small, e.g. by changing the background of the scenario. ii) View changes: highly asymmetric person appearance can lead to a bias for general use cases since every person could appear in any arbitrary rotation variant color appearance, which differs from previously seen samples if open-world data is exploited during inference. Due to the limited number of images in today's public datasets, this claim is not well investigated yet. Instead, increasing key performance indicators suggest tweaking the hyper-parameters of networks [Luo et al., 2019] which, in fact, tend to overfit a particular dataset. Thus, the generalization ability remains questionable due to missing open-world evaluation and a representative number of difficult cases (e.g. asymmetric person appearance or occluded body parts, cf. Fig. 1.2) which is challenging to solve by using such a holistic feature extraction method.

While there are other advanced backbones, backends, and training-strategies which could improve the triplet loss based re-identification performance further (e.g. [Luo et al., 2019]), most of the feature extraction methods share a similar idea and structure mentioned above.

SRNN: The above-introduced methods do not allow the processing of multiple input images of the same person from an image sequence to obtain an embedding, because given one input image, one output embedding is obtained. In this paragraph, we introduce a method that allows us to feed multiple input images of a person into a network to obtain an embedding. This embedding is the result of multiple inputs of the same person in one camera. The used input images can be image sequences to model temporal information or randomly selected images of an image sequence in the hope that a network learns which parts of imaged persons are essential for the re-identification.

In the following section, we review such a strategy which is designed to handle image sequences. We call this method SRNN [McLaughlin et al., 2016] because a *Siamese Recurrent Neural Network*[6] structure is employed, see Figure 3.6. The underlying idea can be described as follows: The network consists of two identical sub-networks with shared weights. Pairs of

[6] For a more detailed theory of Recurrent Neural Networks, the reader is referred to the standard literature such as [Goodfellow et al., 2016].

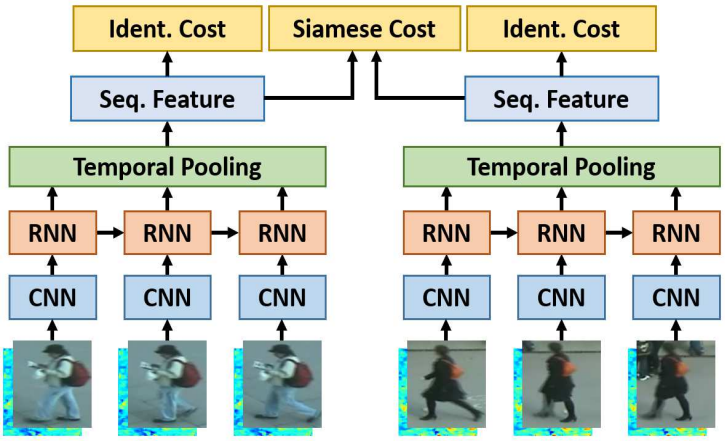


Figure 3.6: General architecture of the SRNN. © Reprinted, with permission, from [McLaughlin et al., 2016].

input-sequences are shown to the network. The sub-networks map the pair of input-sequences to a pair of sequential feature vectors. These are compared, e.g., by using the Euclidean distance. During training, similar and dissimilar input-sequences pairs are shown to the Siamese network, and it learns to map those inputs to a latent space where similar inputs are close, and dissimilar inputs are separated by a margin. During inference, feature vectors are extracted for unseen sequences, and these feature vectors are compared. Consequently, a smaller distance indicates a higher similarity.

In the architectural processing, intermediate features are extracted by a CNN first, and in a further step recurrent neural network layers (RNN) and temporal pooling are employed to obtain a sequence feature vector. This sequential feature vector is created by exploiting spatial image information and temporal information from optical flow. Thus, color is used to model person appearance, and optical flow *"directly encodes short-term motion, which may include details of a person's gait as well as other motion cues." [McLaughlin et al., 2016, p. 1327]*

While the recurrent layers, are able to capture temporal information, they have some drawbacks, those are tackled by McLaughlin et al. by the following thoughts:

"Firstly, the RNN's output may be biased towards later time-steps, making these more dominant than earlier ones. This could reduce the RNN's effectiveness when used to summarise the relevant information over a full sequence, because discriminative frames may appear anywhere in the sequence, not just near the end. Secondly, time-series analysis usually requires extracting information at different time scales. [...] Since multiple time scales are not explicitly encoded in the standard RNN architecture, the temporal hierarchy present in the input signal may need to be explicitly embedded into the network design. In order to address these limitations, our architecture adds a temporal pooling layer. This layer allows for the aggregation of information across all time steps, thus avoiding bias towards later time-steps. The temporal pooling layer aims to capture long-term information present in the sequence, which in combination with the

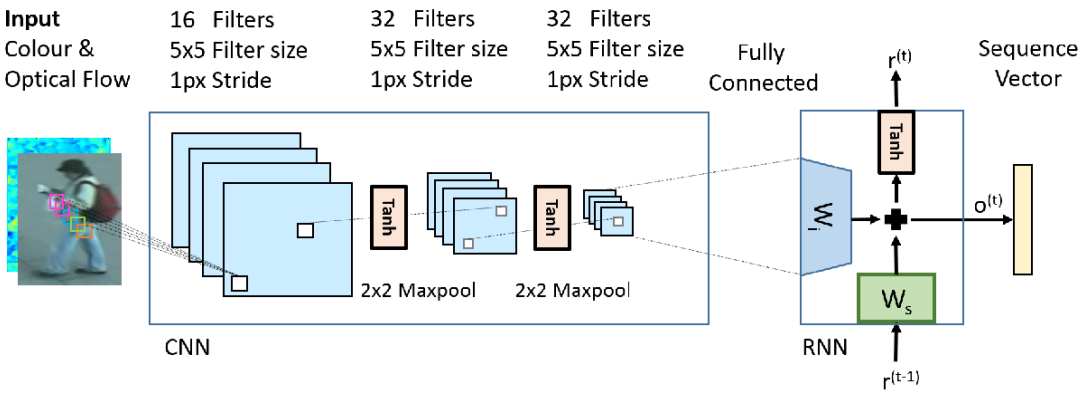


Figure 3.7: Details of the architecture of the SRNN. © Reprinted, with permission, from [McLaughlin et al., 2016].

short term scale of the optical flow input, and the middle-term recurrent layer, aims to model information at all temporal scales within the input signal." [McLaughlin et al., 2016, p. 1328]

In general, a particular backbone and backend are employed, see Figure 3.7. For the network, optical flow from the *Lucas-Kanade* algorithm [Lucas & Kanade, 1981] is coded as horizontal and vertical optical-flow-components, in addition to the three colour channels of an input image. Thus, no standard backbone such as ResNet50 is applied, but an architecture which can handle two additional input channels.

The feature map of the last CNN layer $\mathbf{f}^{(t)} \in \mathbb{R}^{N\times 1}$, from time t is used, together with the recurrent connection from the previous time step $\mathbf{r}^{(t-1)} \in \mathbb{R}^{e\times 1}$ to calculate the output of the RNN as:

$$\mathbf{o}(t) = \mathbf{W}_i f^{(t)} + \mathbf{W}_s r^{(t-1)}, \tag{3.27}$$

where $\mathbf{W}_i \in \mathbb{R}^{e\times N}$ and $\mathbf{W}_s \in \mathbb{R}^{e\times e}$ are fully connected layers, with N is the dimension of the vectorized last CNN feature map, and e is the dimension of the feature space. Moreover, the current RNN state, which is used for the next time step, follows as

$$\mathbf{r}^t = tanh(\mathbf{o}^{(t)}), \tag{3.28}$$

where $tanh$ is the tangens hyperbolicus activation function. Note, in the first step, $r^{(0)}$ is a zero-vector. Moreover, since W_i is non-square the CNN output is mapped to a lower-dimensional feature space. Finally, temporal pooling is used over the T image frames to fuse the output vectors of the RNN to the final sequence feature vector

$$\mathbf{\Phi} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{o}^{(t)}. \tag{3.29}$$

Since a Siamese network is employed, the aforementioned triplet loss function is not applicable. [McLaughlin et al., 2016] argue that for the Siamese architecture *Identity cost* and *Siamese cost* are critical for convergence. The applied loss function $\mathcal{L}$ can be introduced as follows: given an

input-sequence pair ($\mathbf{S}_i = \{\mathbf{I}_{i,t}\}_{t=1,\mathcal{T}_i}$, $\mathbf{S}_j = \{\mathbf{I}_{j,t}\}_{t=1,\mathcal{T}_j}$), the Siamese cost can be calculated by the sequence feature vectors $\mathbf{\Phi}_i = R(\mathbf{S}_i)$ and $\mathbf{\Phi}_j = R(\mathbf{S}_j)$, where $R(\bullet)$ gives the output of the feature sequence extraction network, and

$$\mathcal{E}(\mathbf{\Phi}_i, \mathbf{\Phi}_j) = \begin{cases} \frac{1}{2}||\mathbf{\Phi}_i - \mathbf{\Phi}_j||^2 & i = j \\ \frac{1}{2}max[m - ||\mathbf{\Phi}_i - \mathbf{\Phi}_j||, 0]^2 & i \neq j. \end{cases} \tag{3.30}$$

When $i = j$, the loss function encourages the features $\mathbf{\Phi}_i$ and $\mathbf{\Phi}_j$ to be close, while for $i \neq j$ the function encourages the features to be separated by a margin m. The Identification cost can be calculated by using the standard cross-entropy loss as

$$\mathcal{I}(\Phi) = P(q = c|\mathbf{\Phi}) = \frac{exp(\mathbf{W}_c\mathbf{\Phi})}{\Sigma_k exp(\mathbf{W}_k\mathbf{\Phi})}, \tag{3.31}$$

where there are a total of K identities, q is the identity of the person, and $\mathbf{W}_c$ and $\mathbf{W}_k$ refer to the c^{th} and k^{th} column of $\mathbf{W}$, the softmax weight matrix [Goodfellow et al., 2016], respectively. The overall loss for two sequences ($\mathbf{S}_1$,$\mathbf{S}_2$) is optimized by back-propagation and can be written as

$$\mathcal{L}(\mathbf{S}_1, \mathbf{S}_2) = \mathcal{E}(\mathbf{\Phi}(\mathbf{S}_1), \mathbf{\Phi}(\mathbf{S}_2)) + \mathcal{I}(\mathbf{\Phi}(\mathbf{S}_1)) + \mathcal{I}(\mathbf{\Phi}(\mathbf{S}_2)). \tag{3.32}$$

McLaughlin et al. show that optical flow improves the method by more than 6% rank#1, compared to the experiments where optical flow is not considered. Besides, the experiments not surprisingly indicate that the longer the input image sequences, the better the re-identification performance. For the iLIDS-VID dataset (see Appendix A.2) a performance boost, from using a single image to an image sequence with 128 images, of +38 rank#1 is found (14% to 52% rank#1), whereas for using an image sequence of 16 images 36% rank#1 is obtained. We believe the reason, for better performance by using more frames, is that more diverse frames are needed to improve the performance, whereas using a higher frame rate to obtain more images only helps to see the same person multiple times from the same view without new information about this person, i.e. the information content does not increase.

Chapter 4

A new approach for person re-identification

In this chapter, we present our novel approach. First, a general overview is presented in Section 4.1. Section 4.2 to Section 4.6 then present a detailed discussion of individual components. Finally, Section 4.7 closes the chapter with a discussion of the characteristics of the whole approach.

4.1 General overview

Despite tremendous progress achieved by learning-based methods, the re-id performance is still not sufficient for autonomous person re-identification scenarios. One major issue is the intra-person and inter-person variation problem, which is discussed in Section 1.3. Aiming to tackle the above problem, in this chapter, we present a novel approach to re-identifying individuals with multi-view observations.

The high intra-person variation and low inter-person variation problem can also be stated as a problem of limited observations to distinguish between individuals. That means fine-grained glimpses and a complementary person view are usually not available to uniquely describe and differentiate between persons.

Selecting a standard camera with a central projection lens and wall, ceiling, or egocentric suspension complicates the situation due to the sensor principle in combination with the underlying optical path: Persons typically cannot be captured in various different views to obtain the mandatory variety of observations in the respective camera field of view to solve the appearance-based person re-identification. Examples of exceptions are discussed later.

Imagine a person with a high intra-person variation, e.g. highly asymmetric appearance around the longitudinal axis (cf. Fig. 4.1). Analyzing this person in a first camera to re-identify the same in a second camera, where the person is only observable in an opposite view, can be a challenging task. Other factors affecting the re-identification performance are the gallery size and the diversity of persons in the gallery. However, in the open-world highly asymmetric person appearance is one of the essential issues (see Fig. 4.1 for example) in:

- Indoor environments, such as shops, where persons tend to wear open jackets which significantly differ from the appearance of underlying clothes.
- Scenarios consisting of highly fashion-conscious persons.
- Scenarios where persons are partly occluded, e.g. by huge backpacks, suitcases, shopping carts, and strollers which are not detected and modeled individually in the re-id stages, but rather directly integrated into the appearance representation.

To allow for a large field of view for multi-view person observations in a single camera, we modify the optical path and use a fisheye camera system. Thus, our approach consists of both: i) a hardware setup to allow obtaining multi-view observations, and ii) a novel analysis approach to process multiple observations of persons.

Our approach is designed to systematically provide and handle different views per individual in each respective camera. Rather than view specific PRID methods (cf. Sec. 2.6), where different view combinations are matched with individually learned weights, we suggest a more general *multi-view* approach: i) different views per person are provided and ii) treated independently of each other instead of fitting weights for cross-view matching to a specific dataset of arbitrary view combinations, which has restricted statistics and can lead to a bias for general use cases.

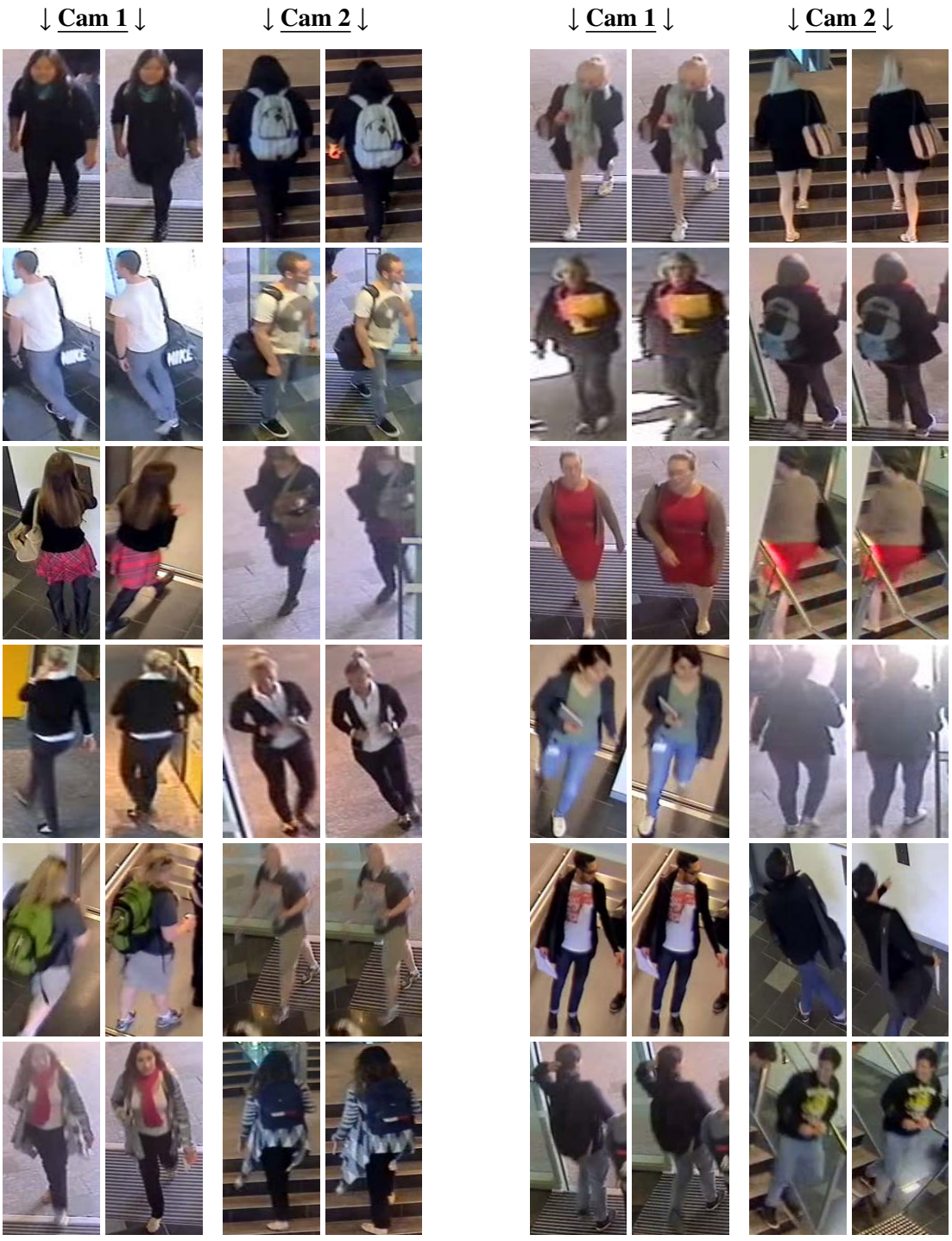


Figure 4.1: Illustration of high intra-person variation: 12 individuals are shown, who are observed in two cameras of disjoint fields of view. Two images per person per camera are depicted, respectively. Persons are picked for illustration from the SAIVT-BuildingMonitoring database [Denman et al., 2015], which consists of a recorded open-world campus scenario. Due to a different focus, this dataset is not annotated for person re-identification with global id association. However, as the illustrated persons indicate, highly asymmetric appearance is present and influences the re-identification, in particular, if persons appear in another camera in different appearance due to the view. We thank the SAIVT Research Labs at Queensland University of Technology for supplying us with the raw-videos.

Hardware setup

To allow for multi-view person re-identification between single cameras within a monocular camera network of disjoint fields of view, we propose the employment of a fisheye (FE) camera in nadir pose (cf. Fig. 4.2).

Fisheye cameras are less popular than central projection cameras in the security camera domain. However, in a nadir pose, they enable *within-video multi-view information*[1] in each single camera. Multi-view information can help to obtain and classify different views of persons. This allows the treatment of the views independently of each other during the matching process, since, in good-natured cases, persons are available in a surround-view.

Specifically, the special property of the FE lens mounted in nadir direction is used, to provide different views per person in one camera: a person directly under the camera is recorded in an image showing only the head, whereas with increasing distance to the optical axis (off-axis displacement) the body of the person becomes more and more visible. Figure 4.2 illustrates the situation in a realistic scenario. Whereas usually a camera using central projection (Figure 4.2 (a) left) provides one view per person per camera only, the FE camera is able to provide several different views if persons are moving in the scene.

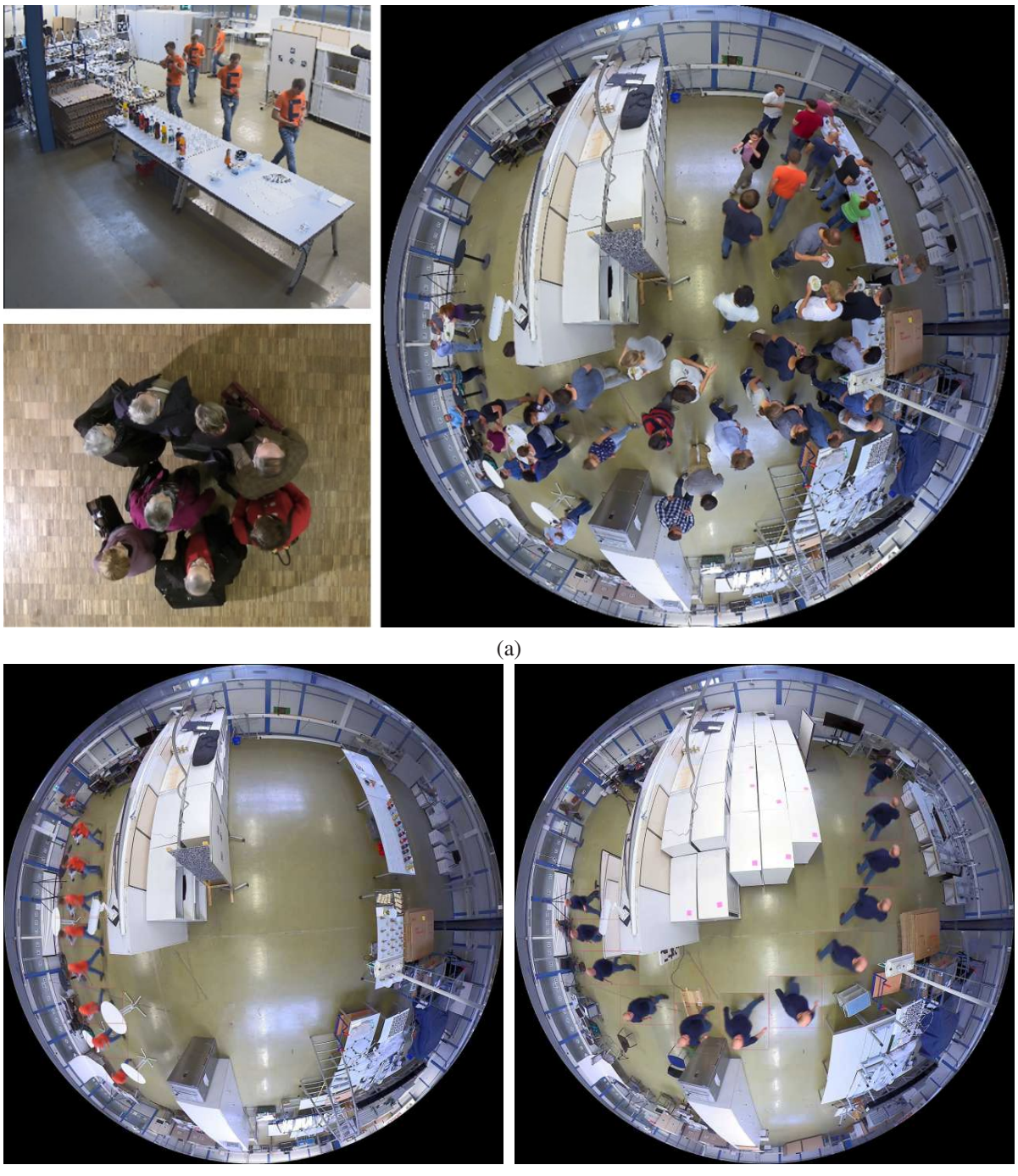
In general, by appropriately guiding persons - e.g. by tapes on the ground or by furniture (cf. Fig. 4.3) – also a central projection camera can provide different views – if the persons follow the pre-defined trajectory. However, this limits the use cases. Nonetheless, a simple modification of our software pipeline allows us to use it for this special purpose as well.

In benefitting from the nadir camera pose, possible scenarios are limited to ceiling suspension, which is usually applied in stores, shops, airports, railway stations, and fairgrounds. Here controlled scene illumination, which is beneficial for PRID, is typically mandatory. Compared to a wall suspension, person occlusions are minimized due to the particular camera pose.

Software pipeline

To handle the highly variable within-video multi-view information acquired by the fisheye cameras, we propose a modular processing pipeline, which is illustrated in Figure 4.4 and briefly presented in the following section.

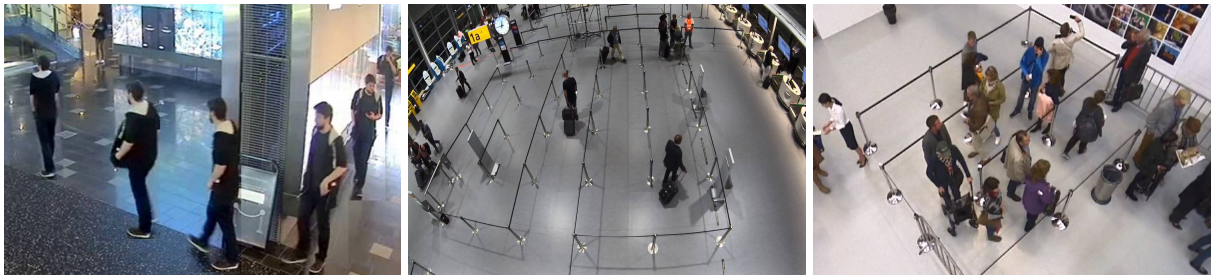
[1] This can also be called *intra-video multi-view information*.



(a)

(b) (c)

Figure 4.2: Motivation for the use of a fisheye lens for PRID. **(a)** Left (central projection): narrow field of view, random camera pose (top), and nadir pose of another scene (bottom). Right (fisheye projection): nadir camera pose. **(b)** Person from (a) (top left) follows a straight path and can be observed in front, back, and side views. As can be seen, the fisheye image contains different views of a person if he/she moves in the scene. It is nearly impossible to obtain a similar result with central projection cameras. **(c)** Another person follows a random trajectory, much more detailed view information can be observed, e.g. bird's eye views, which show fine-grained head observations in high resolution. Note that ((a) top,left), (b), (c) are multi-exposure images (video synopsis) to demonstrate the content of a video in a single image in this figure. Later in Section 5.5.3 a more detailed fisheye image illustration is given in Fig. 5.12.



(a) (b) (c)

Figure 4.3: Examples of guiding persons in scenarios which allow the obtainment of multi-view information: **(a)** Multiple-exposure image (video synopsis) of a person following a trajectory around the wall in the building. Image created based on the raw material of [Denman et al., 2015]. **(b)+(c)** Persons are guided with border tape in a queue, e.g. an airport or waiting queue. Applying within-video tracking allows the analysis of persons in different views, however, this setup is a restriction compared to our setup.

Overview (cf. Figure 4.4):

(1) Input: Consider an unknown person (probe) observed in one camera. Then, the input to our pipeline is a circular fisheye image sequence ($\{\mathbf{I}_{\mathcal{P},t}\}_{t=1,\xi_{\mathcal{P}}}$) consisting of the probe person ($\mathcal{P}$), with $\xi_{\mathcal{P}} \in \mathbb{N}_+$ image frames. The sequence is pre-processed by a frame-wise person detection and tracking which is not the focus of this work.

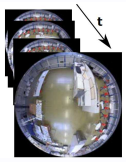
(2) Projection alignment: In fisheye-image-space, the shape of a person standing on the ground plane depends on the position in object space relative to the camera, and the camera elevation above ground. Moreover, persons are aligned quite similarly to a watch hand centered around the principal point. The goal of the *projection alignment* step is to align persons upright (vertical body axis along the 12 o'clock hand) and image filling into a *virtual central projection pinhole camera* image. The benefit is i) a dominant person orientation, ii) images without any distortions, and iii) a fixed image size[2] which is exactly the network input for the further steps. Therefore, any training data and feature extraction method from the central projection domain can be used, without the need to tailor them to a proper fisheye image solution. Thus, no fisheye domain-specific training is necessary. Rather, data and a network architecture of the central projection domain can be shared, whereas the benefit of using a fisheye lens to obtain multi-view observations is preserved.

(3) View classification and sampling: The resulting images of (2) show a person from different views. A deep learning based view classifier is applied to distinguish three different view classes: *front*, *back*, and *side*. The fourth class, *bird's eye views* (BEV), can easily be derived from the position in image-space. Person re-identification with bird's eye view images shows increased ambiguities in appearance-based re-identification. Therefore, we omit bird's eye views for person re-identification (see also the discussion in Section 5.3).

[2] $\zeta = w \times h \times c$; virtual image width ($w \in \mathbb{N}_+$), virtual image height ($h \in \mathbb{N}_+$), and number of color channels (here, $c = 3$).

(1) **Input**

A fisheye (FE) image sequence consisting of an unknown person (probe) is used as input. The sequence is pre-proccessed by a frame-wise person detection and tracking. (image details best seen in Figure 4.2 (b) + (c))



(2) **Projection alignment**

Obtain images of the person in dominant orientation and as taken by a central projection camera. Afterward, a deep learning network architecture and weights, which are initalized on public datasets, can be used without a particular FE adaptation.

do: i) Align the probe from the FE images to the dominant orientation. ii) Eliminate FE distortions to reduce ambiguities. iii) Obtain a fixed output size. Output image is obatained by single indirect pixel look-up in the input image.

(3) **View classification and sampling**

Obtain for each image of (2) the view of a person and use the view as strong prior knowledge in consecutive steps to reduce ambuguities during matching.

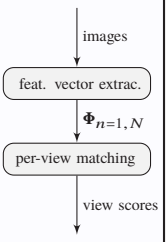
do: i) Classify the view. ii) Sample a 3-view set consisting of best front, back, and side views. iii) The fourth class "bird's eye view" is obtained by the position in the image to omit bird's eye view images from further processing.

Gallery Images

(4) **Per-view matching**

Same views are compared to reduce ambiguities.

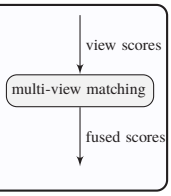
do: i) In advance to the probe processing, steps (1)-(3) are applied to the persons in the gallery. Note, step (2) is only needed for fisheye images. ii) The feature extraction method is applied. iii) Same views are matched. In case an individual is only available in one view, per-view matching is reduced for this person to an available-view-matching.



(5) **Fusion**

Obtain a matching-score which is based on a consensus decision of per-view matches.

do: i) Fuse with respect to the views, and ii) apply a multi-view matching.



(6) **Output**:

Locate the most probable multi-view match between probe and gallery, which shows highest similarity across different views.

do: Find lowest score which indicates highest similarity.

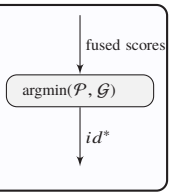


Figure 4.4: Overview of the proposed PRID processing pipeline

Best shots are then selected from the classified images in the *sampling* step. As a result, for each probe person we obtain a 3-view set $\mathbb{S}_{\mathcal{P}} = [\mathbf{F}^{\zeta \times i}, \mathbf{B}^{\zeta \times j}, \mathbf{S}^{\zeta \times k}]$ consisting of a $\zeta \times i$ tensor of front views (**F**), a $\zeta \times j$ tensor of back views (**B**), and a $\zeta \times k$ tensor of side views (**S**).[3] Generally speaking, the tensors are introduced to provide a structure where same views are stored.

(4) Per-view matching: The major goal of this stage is a matching of corresponding views. Matching of the same views helps to decrease the intra-person variation and increase inter-person variation, since the number of ambiguities is decreased. To do so, this stage uses a feature extraction method and applies the per-view matching to the *3-view sets* of (3).

Persons previously observed in other cameras build the gallery $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_N\}$. Thus, step (1) to (3) of the pipeline are applied again for all images in the gallery to obtain 3-view sets ($\mathbb{S}_{\mathcal{G}}$) and embeddings.[4]

Since the number of related work dealing with new features is rapidly increasing we decided to model the feature extraction in our pipeline in a more general manner. Thus, different feature extraction methods are applied as a module and evaluated in the experimental part of this work to show the general improvement of multi-view compared to strategies such as single-shot and multi-shot or view-aware matching and random view matching.

In the case that only views of different directions are available, to match a probe with a particular person from the gallery, our multi-view approach is then, for this particular combination, reduced to a multi-shot matching. One example is the case where a probe is only visible in a front view and side view, whereas a candidate from the gallery is only available in a back view.

(5) Fusion: A fusion is applied to obtain a consensus decision per candidate, i.e. results of per-view matching are considered to use this strong prior knowledge to obtain one similarity score for each possible candidate, which is based on multiple per-view matchings. The matching compares the distances between the probe person ($\mathcal{P}$) and all candidates in the gallery ($\mathcal{G}$) to find the smallest distance i^* which implies the most probable match,

$$i^* = \ argmin_{i \in (1,2,...N)} ||f(\mathcal{P}) - f(\mathcal{G}_i)||_2. \tag{4.1}$$

An image of $\mathcal{P}$ in the probe camera and an image of $\mathcal{P}$ in the gallery camera is a mandatory prerequisite to find the corresponding match. As discussed in Section 2.3, this process is called the "closed set scenario" as that particular person was seen before, and it is certain that a corresponding match is provided in the gallery.

(6) Output: The output of the pipeline is the match of the smallest distance i^*, which is taken as the pair with the most similar persons. Re-ranking methods, as discussed in Section 2.7 are not applied in our work, since they can typically only improve the performance further if

[3] Note, $i + j + k <= \xi_{\mathcal{P}}$ and $i, j, k \in \mathbb{N}_0$

[4] Concerning an efficient implementation, this step is carried out before an unknown probe is available.

multiple samples are provided, e.g. by recording with higher frame rates at the expense of higher computational costs and shorter exposure time. Furthermore, state-of-the-art re-ranking methods are designed to find similar appearance-matches of the same person in the gallery to improve the certainty of matching, whereas we pursue the goal of obtaining different views to handle highly asymmetric appearance, which by definition of "asymmetric" does not share the same appearance in multiple views.

In the following section, the individual processing stages are discussed in detail.

4.2 Input and assumptions

In this work, we focus on analyzing multi-view person images from fisheye cameras in nadir viewing direction in a multi-camera network and integrate view information as essential prior knowledge into our pipeline (cf. Fig. 4.4).

In first unpublished experiments it became apparent that state-of-the-art deep learning methods based on central perspective images acquired with a horizontal optical axis, such as in public datasets like MS COCO [Lin et al., 2014], ImageNet [Deng et al., 2009], Cityscapes [Cordts et al., 2016], or the Multiple Object Tracking Benchmark [Leal-Taixé et al., 2015, Milan et al., 2016], might fail in detecting and tracking of persons, due to the strong fisheye distortions and a missing dominant person orientation in the images from a nadir viewpoint.[5]

We thus assume person detection and tracking in this work to be given, as we want to concentrate on the aspect of re-identification, and therefore developing a sound solution for detection and tracking in the highly distorted fisheye images is beyond the scope of this thesis.[6] Thus, in this work all the outputs of detection and tracking are obtained by manual annotations of the dataset. One naive way to obtain the required input automatically is to warp the fisheye images to perspective geometry (cf. Fig. 4.5) and to then apply state-of-the-art detection and tracking trained with publicly available datasets. A drawback of this approach is that due to the camera orientation *perspective distortions* can appear, some image information is lost during the rectification process, and spatial image resolution decreases with increasing radius from the principal point in the fisheye image. Thus, persons in the de-warped images show decreasing quality from the shoes to the head. This is not the case for a real central projection camera with

[5] The issue of training a deep learning based architecture without camera-pose-specific training material can be underlined by our work [Blott et al., 2018a] where we propose a novel approach to improve the semantic segmentation of fisheye images without using any fisheye training data. The results of conducted experiments indicate that our proposed method improves the key performance indicators by a noticeable margin. However, to bridge the domain gap between egocentric camera pose and security camera pose a vast amount of central projection images from the nadir and oblique camera poses are used, and only a moderate segmentation performance was obtained which is far from sufficient for this work.

[6] Some methods regarding person detection and person tracking in fisheye images, from the pre-deep-learning-era, are presented in [Ibrahim, 2011] [Demiröz et al., 2012] [Imran, 2014], [Chiang & Wang, 2014].

Figure 4.5: Illustration of a general de-warping result where one fisheye image was de-warped to four central projection images. Left: fisheye reference frame. Right: reference frame is de-warped to four central projection images (south-west, north-west, north-east, south-east).

tripod or wall suspension. Thus, resulting changes in image statistics will need an adaptation of networks for this domain.

Consequently, the chosen decision, to consider person detection and tracking as given, allows the determination of an upper bound of performance, which will not normally be reached in practical application due to non-perfect person detection and tracking. In the experimental part of this thesis, we simulate the effect of tracking errors to study their influence on PRID (cf. Sec. 5.5.5).

4.3 Projection alignment

A significant drawback of an FE camera is the fact that, due to the lens properties, persons are often heavily distorted. The amount of distortion increases with increasing distance to the principal point: for persons next to the optical axis only the head is visible, but with negligible distortions, while persons next to the image border suffer much larger distortions.

The most important strength of our fisheye cameras is that it they allow the observation of multi-view information from monocular cameras. However, this information cannot be obtained all over the image. In the fisheye image space, regions can be categorized into four classes which are depicted in Figure 4.6.

$\mathcal{U}$ is the passepartout of the image caused by the projection of light through the lens and aperture onto the sensor. This region does not contribute to any observation of the scene.

$\mathcal{B}$ presents the central image region where persons are only visible in bird's eye views. The fisheye effect can be neglected but ambiguities in person-appearance increase due to the camera pose.

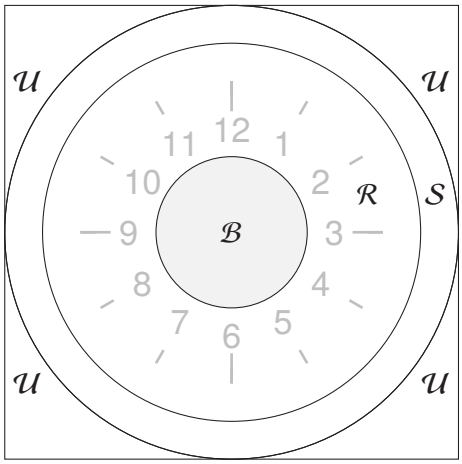


Figure 4.6: Schematic illustration of available regions within a nadir fisheye camera image. $\mathcal{U}$ s show the image passepartout, $\mathcal{B}$ the bird's eye view region, $\mathcal{S}$ the region where persons are squeezed too much by the projection, and $\mathcal{R}$ indicates the region which is preferred for person re-identification. Further, in $\mathcal{R}$ the dial illustrates some of the orientations, in which persons appear.

$\mathcal{S}$ shows the region where persons are heavily distorted and squeezed into tiny patches. Thus, it can only marginally contribute to PRID since up-sampling to larger image resolution will introduce many interpolated pixels which do not necessarily describe the real person's appearance.

$\mathcal{R}$ depicts the region where persons can be analyzed in different views when they move in the scene. This is the preferred area for the person re-identification task.

As a result, only persons in $\mathcal{R}$ are considered for further processing.

To reduce the effects of distortion, after person detection and intra-camera tracking over time in the FE images, de-warping and alignment of the person to a rectilinear image as taken by central projection from a virtual pinhole camera are applied; in other words, persons are mapped to a dominant pose independent of their initial position in the FE image.

Certainly, the resulting quality and shape of person images deteriorates with increasing off-axis displacement. Shapes in fisheye image space are distorted and rotated depending on their position in object space relative to the optical axis (cf. Fig. 4.2). For PRID, however, it is advantageous (and for many algorithms mandatory) to have images of constant size and shape with persons in an upright position as input. Furthermore, most deep learning based feature extraction methods used for PRID sample all input images to the same image size in a first step to obtain the network input. A transformation from FE space to that of the virtual pinhole camera, called vCam in the following, is thus needed. This transformation, called projection alignment, eliminates lens distortions, changes the scale of the bounding box containing the person, and rotates it to show the person in an upright position. The interior orientation of the FE camera is assumed to be known to carry out this transformation; it can be determined in a prior calibration step [Strauß et al., 2014] using *Mei's* projection model (cf. Sec. 3.1). Note that photometric

camera calibration, e.g. the handling of vignetting effects and light falloff, is not considered here, as this is assumed to be corrected by the camera device in a pre-processing step.

The projection alignment can then be expressed as

$$\mathbf{x}_{FE} = \mathbf{P} \cdot \mathbf{x}_{vCam}, \tag{4.2}$$

where $\mathbf{P}$ is the partly unknown projection matrix and $[\mathbf{x}_{vCam}, \mathbf{x}_{FE}]$ are the coordinates of corresponding points in the respective image spaces. Further, $\mathbf{P}$ can be decomposed into a $\mathbf{P}_{OS}$ part for orientation and scaling, and a warping part $\mathbf{P}_{warp}$:

$$\mathbf{P} = \mathbf{P}_{warp} \cdot \mathbf{P}_{OS}(\alpha, \beta, c_{vCam}), \tag{4.3}$$

we will discuss the rotation parameters (α, β) and the focal lenght of the virtual camera (c_{vCam}) later on.

For $\mathbf{P}_{warp}$, the projection model introduced by Mei [Mei, 2007][7] is used, which is an extension of [Barreto & Araújo, 2001, Geyer & Daniilidis, 2000]. The fundamentals are described in Section 3.1.

To generate the vCam image from the FE image we start with a point in vCam space and apply the model of [Mei, 2007] to obtain points (X_S, Y_S, Z_S) on the unit sphere, and then project these points into the FE camera (cf. Fig. 4.7) where bilinear interpolation is applied to obtain the vCam image pixel values.

As described, the interior orientation of the FE camera is assumed to be known (we use the one determined based on [Strauß et al., 2014]). Thus, all elements of $\mathbf{P}_{warp}$ are given.

In contrast to $\mathbf{P}_{warp}$, the transformation parameters of $\mathbf{P}_{OS}$ are unknown and depend on the person location in object space. Using

$$\mathbf{P}_{OS} = \mathbf{R}(\alpha, 0, \beta) \cdot \mathbf{K}_{vCam}^{-1}(c_{vCam}, 0, 0), \tag{4.4}$$

with $\mathbf{R}$ a SO(3)[8] rotation using two Euler angles (cf. $[\alpha, \beta]$ in Fig. 4.7), and $\mathbf{K}_{vCam}$ the camera calibration matrix, which includes focal length and principal point.

The three unknowns α, β, and c_{vCam} can be solved one after the other. Again, the goal is to align the person without distortions in a vCam image where the person's head and footpoint lie at the top and bottom image borders, respectively, and all images have the same number of pixels. The unknowns can be solved in three steps: the centre coordinates of the bounding box (Bb) in the FE image are used to determined the rotation angle $\alpha = \arctan(x_{Bb}, y_{Bb})$ (cf. Fig. 4.7). β is selected such that the principal point of the vCam becomes the center between head and

[7] http://www.robots.ox.ac.uk/~cmei/articles/projection_model.pdf
[8] Special orthogonal group of dimension three

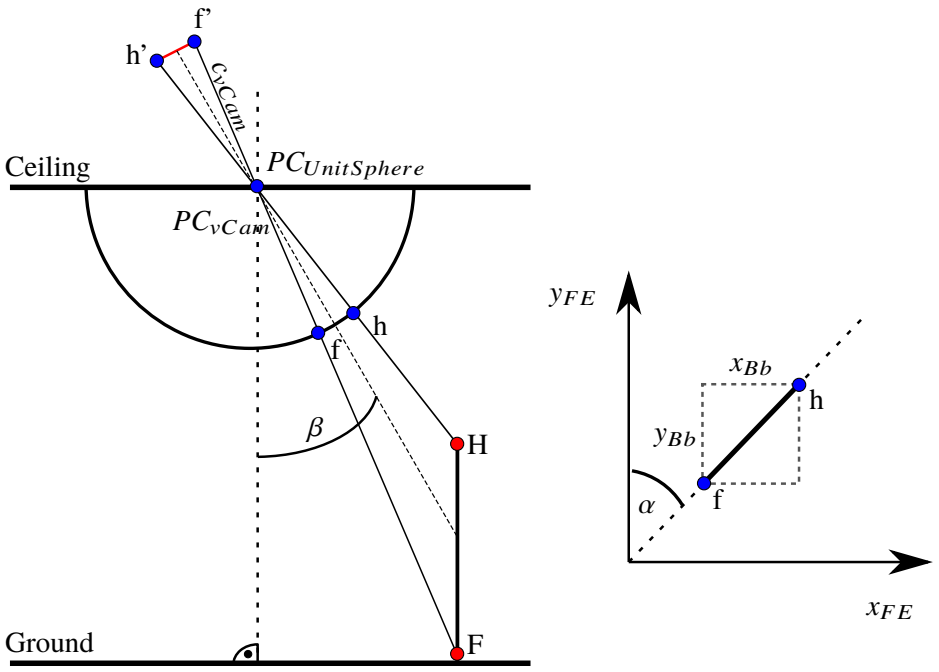


Figure 4.7: Illustration of projection alignment. Left: points in object space (foot and head points $[F,H]$) are projected onto a unit sphere as in [Mei, 2007], $[h,f]$ show the results. Then, the further steps of Mei are applied to obtain FE coordinates (not shown here). Using the virtual camera with the unit sphere origin as the projection center allows projecting points into the virtual image. Thus, once FE image points are back-projected to the unit sphere, they can be mapped to the virtual rectilinear image. In this way, $[h,f]$ are mapped to $[h',f']$. Right: foot and head points $[h,f]$ in FE camera space, a person is rotated around α. $[y_{Bb},x_{Bb}]$ are the height and width of a bounding box. The projection center of the unit sphere ($PC_{UnitSphere}$) and the projection center of the virtual camera (PC_{vCam}) are identical.

footpoint (approximated person center). Finally, c_{vCam} is determined so that the person height in pixels (head point - foot point) is scaled to the pre-defined target image size in pixels (we use images of 128 pixels in width and 256 pixels in height). Thus, by using

$$\mathbf{K}_{vCam} = \begin{pmatrix} c_{vCam} & 0 & \frac{128}{2} \\ 0 & c_{vCam} & \frac{256}{2} \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.5}$$

c_{vCam} can be determined by a reformulation of the following equation system, since there are no other unknowns remaining:

$$\mathbf{x}_{FE_{foot}} = \mathbf{P}_{warp} \cdot \mathbf{R}(\alpha, 0, \beta) \cdot \mathbf{K}_{vCam}^{-1}(c_{vCam}) \cdot \mathbf{x}_{vCam_{foot}}. \tag{4.6}$$

Note, for $\mathbf{x}_{FE_{foot}}$ the closest point within the bounding box with respect to the principal point is assumed to represent the foot point of a person. For a more precise detection of keypoints semantic segmentation or the person pose is needed. After applying the resulting transformation step, cropped person images look as if they were taken by a central projection camera and any deep learning based method can be applied without the need to tailor them to a specific fisheye-distortion architecture.

4.4 View classification and sampling

In this stage, views are classified and best samples are selected. As de-warped images are used as input, existing perspective image datasets and annotations can be used to pre-train a network, such as with the ImageNet dataset [Deng et al., 2009]. This stage can be divided into a view classification and a sampling step.

View classification: We classify four different view classes: *front*, *back*, and *side*, whereas the fourth class, *bird's eye views* (BEV), can easily be derived from the position in image-space. We motivate the design decision that we use three horizontal views as follows: i) The effect of matching with different views is studied recently in [Sun & Zheng, 2019], where a large scale synthetic dataset is shown for person re-identification, and all individuals are available in different cameras and views (in [Sun & Zheng, 2019] called rotation angles; $\measuredangle = [0°, 360°]$ with $10°$ increments). Conducted experiments indicate that matching left with right views changes the performance only slightly, whereas matching front and back views deteriorates the performance by a large margin. The two classes front and back are thus necessary. Also, a person passing the fisheye field of view is typically only visible in left or right viewing-angle-cluster. While it is true that *left* and *right* might have a different appearance, we combine these two classes to simplify and robustify further processing. Moreover, the results of [Sun & Zheng, 2019] suggest that there will hardly be a performance drop. ii) Person re-identification with finer-grained viewing angles needs a robust classification of those angles, but to the best of our knowledge, there is hardly any work dealing with this task. Since we do not have access to datasets to investigate a large number of views, we restrict ourselves to the three mentioned views.

Finally, we note that only when using horizontal views (rather than, e.g. oblique views), we can use publicly available datasets and annotations for fine-tuning a pre-trained network with datasets such as Market-1501 [Zheng et al., 2015a].

Detections of a person in a viewing angle of less than around $20°$ between bounding box centre and the principal point in the fisheye image, are considered as bird's eye views. The use of bird's eye view images typically shows increased ambiguities in appearance based re-identification, since people seen from above can be hard to distinguish. Avoiding such ambiguities, these bird's eye views are eliminated from further computations.

To classify the single person images cropped from the FE images into "front", "back" and "side", three popular deep learning architectures are evaluated as a backbone. i) InceptionV3 [Szegedy et al., 2016], ii) ResNet50 [He et al., 2016], and iii) VGG16 [Simonyan & Zisserman, 2014]. All networks are pre-trained on the ImageNet dataset [Deng et al., 2009]. The last fully connected layer(s) of all backbones are replaced by an average-pooling layer, one 1024 dimension fully connected layer with *rectified linear unit* (ReLU) activation and a fully connected layer with *softmax* activation [Bishop, 2007], see Figure 4.8.

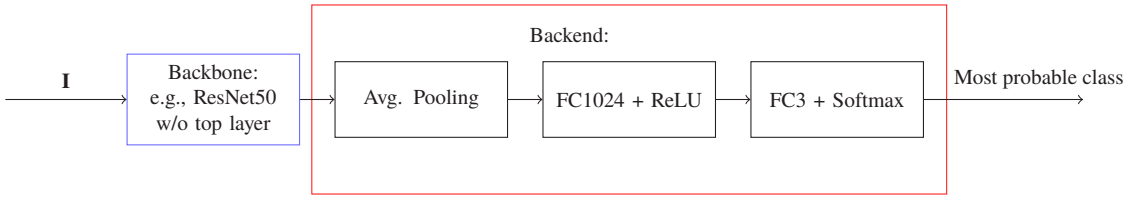


Figure 4.8: Deep Learning Architecture used for the view classification

Note, we have decided against a view classification using person tracking (cf. Sec. 2.6), since such approaches share the assumption that persons only move in the forward-moving-direction. Thus, in contrast to these approaches, our view classification can handle arbitrary moving directions of persons.

Sampling: As a result of view classification, we obtain a 3-view set $\mathbb{S}_{\mathcal{P}} = [\mathbf{F}^{\zeta \times i}, \mathbf{B}^{\zeta \times j}, \mathbf{S}^{\zeta \times k}]$ for each probe and gallery person, consisting of a $\zeta \times i$ tensor of front views ($\mathbf{F}$), a $\zeta \times j$ tensor of back views ($\mathbf{B}$), and a $\zeta \times k$ tensor of side views ($\mathbf{S}$).[2] At the same time, within the tensors, the cropped person images are ordered by decreasing confidence.

To generate 3-view sets in the sampling stage, the best views of each class are selected based on the highest confidence of the corresponding view classifiers, which is the highest softmax activation for the respective class over all images of one person in the respective camera.

While we could also take multiple images per class ($i, j, k > 1$), we do not have such an abundance of data in our experiments. Therefore, we set $i = j = k = 1$ in this work which we discuss in Section 5.5.

4.5 Per-view matching

The goal of this stage is i) to describe imaged persons in the latent space (feature extraction), and ii) then to apply a per-view matching to the *3-view sets*.

Feature extraction: Many methods exist to extract features from cropped person images (cf. Chapter 2). Our modular pipeline allows us to use an arbitrary feature extraction method as long as a d-dimensional representation is obtained. Therefore, we apply and evaluate different state-of-the-art methods from related work, including handcrafted and deeply learned features. In particular, we use:

- *GOG* for feature description [Matsukawa et al., 2016] and *XQDA* [Liao et al., 2015] for metric learning (cf. Sec. 3.2.1): This method is a handcrafted feature extraction method, which to the best of our knowledge, shows the highest performance on the small and challenging VIPeR dataset [Gray et al., 2007].

We exploit this feature extraction method since we also use a challenging dataset of small training size later, which is not sufficiently large for deep learning methods.

- *TriNet* [Hermans et al., 2017] (cf. Sec. 3.2.2): This is a state-of-the-art deep learning based holistic feature extraction method using a particular triplet loss function during training. The batch hard loss allows for the mining of hard negatives (a different person with a very similar appearance), and hard positives (the same person with slightly different appearance).

  Whereas other recently published deep learning based feature extraction methods, e.g. [Luo et al., 2019], show in fact higher performance on particular public datasets, we obtained with such methods only moderate performance on internal data of real security cameras compared to the TriNet. This could be explained as overfitting of the architecture to the public dataset statistics or as a limited amount of training samples for the internal data. Since our novel datasets are taken by real security cameras (cf. Sec. 5), we prefer the TriNet.

- Siamese+RNN [McLaughlin et al., 2016] (SRNN, cf. Sec. 3.2.2): This is another deep learning based feature extraction method with recurrent CNN architecture to model temporal context. In contrast to both other methods, the Siamese+RNN architecture proved to be an efficient multi-shot fusion method by aggregating features of images acquired one after another. To handle temporal context, optical flow is exploited. We use this method to quantitatively study the behavior of this method by using multiple views instead of consecutive images.

Per-View matching: There are different possibilities of how to perform a per-view matching and a subsequent fusion. Depending on the details, both components may overlap to some degree, and we, therefore, present the per-view matching with respect to the multi-view matching in the following section.

4.6 Fusion

A major advantage of the proposed approach is the increased number of person views, including the corresponding view-class. After per-view matching the per-view matching results are fused, to obtain a score that describes the similarity between one probe and one person from the gallery in a fused-space.

For multi-view PRID, five fusion methods are proposed in the following which are evaluated in Section 5.2 to find the best method: one early fusion method, three late fusion methods, and one deep learning based fusion method.

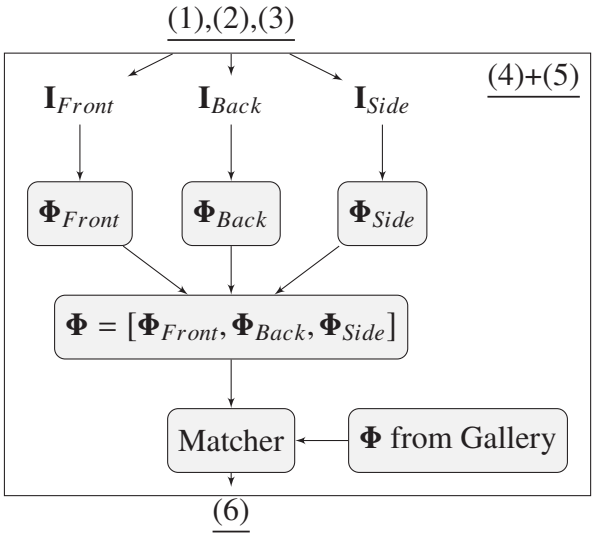


Figure 4.9: Early Fusion: steps (4)+(5) of Figure 4.4 are shown in detail. Fusion with feature vector concatenation.

For early fusion, features extracted from images are concatenated first, after which the resulting feature vector is used for matching. For late fusion, matching of feature vectors is performed separately for each person view, after which fusion is applied to the results. In contrast to the rule-based early and late fusion, deep fusion allows for a dataset dependent learning. In the following section, the different methods are explained in detail; the principle data flow is illustrated in Figure 4.9- Figure 4.11.

Early fusion: feature concatenation

The first fusion method involves a concatenation of feature vectors that are previously extracted from the 3-view sets (cf. Fig 4.9).

The feature vector for one person in one camera can be written as

$$\mathbf{\Phi}^{(1\times(a+b+c))} = [\mathbf{\Phi}_{Front}^{(1\times a)}, \mathbf{\Phi}_{Side}^{(1\times b)}, \mathbf{\Phi}_{Back}^{(1\times c)}],$$

where $\mathbf{\Phi}_{Front} = f(\mathbf{I}_{Front})$, $\mathbf{\Phi}_{Back} = f(\mathbf{I}_{Back})$, and $\mathbf{\Phi}_{Side} = f(\mathbf{I}_{Side})$ and a, b, c are the dimensions of the view-depending feature vectors.

In this fusion method, the per-view matching is applied by i) the pre-defined order of view-specific feature vectors in the concatenated feature vector, and ii) by comparing the distance between two feature vectors.

Late fusion: inverse rank position algorithm (IRPA)

IRPA is an algorithm for merging multiple feature similarity lists into a single overall similarity ranking list, which was proposed for database retrieval [Jović et al., 2006]. After an initial view-

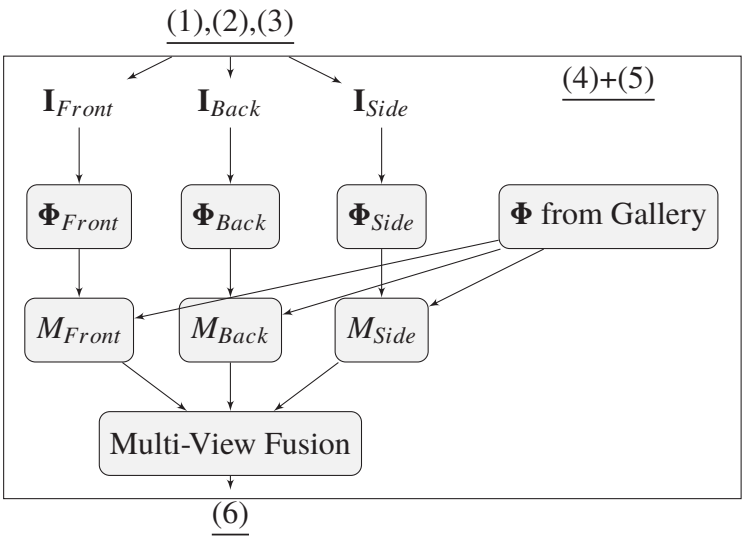


Figure 4.10: Late Fusion (rule-based): steps (4)+(5) of Figure 4.4 are shown in detail. M_μ are the per-view matchings. For *multi-view fusion* three different rules are presented in the coarse of this work.

specific-matching, the fusion is performed (cf. Fig 4.10). The ranking results of front–front, back–back, and side–side matching for all person IDs are transformed into a new ranking result as follows:

$$r(\mathcal{P},\mathcal{G}_i) = \frac{n}{\frac{1}{r_{Front}(\mathcal{P}_{Front},\mathcal{G}_{i_{Front}})} + \frac{1}{r_{Back}(\mathcal{P}_{Back},\mathcal{G}_{i_{Back}})} + \frac{1}{r_{Side}(\mathcal{P}_{Side},\mathcal{G}_{i_{Side}})}},$$

where $\mathcal{P}$ is the queried probe person, $\mathcal{G}_i$ a currently compared person from the gallery, n is the number of fused views (here three), and r the corresponding rank. $[\xi_{Front},\xi_{Back},\xi_{Side}]$, with $\xi \in \mathcal{P},\mathcal{G}_i$, are images of particular views found by the view classification (cf. Sec. 4.4). The final ranks are estimated according to the fused and re-ranked ranking list.

In this fusion method, the per-view matching is applied by directly matching the same views between probe and gallery to obtain a view specific ranking. Results of view specific rankings are then fused. Note, the goal behind *re-ranking* is different, than that in Section 2.7. We intend to apply a first per-view ranking to obtain a rank that describes the similarity between a probe and a person from the gallery in comparison to other person from the gallery. Then, we fuse these per-view ranking results to a per person ranking result. In contrast to this goal, the related work does not use any view information nor handles asymmetric appearance, instead, multiple samples of the same person are assumed to share the same appearance, and, thus, ranking-lists are re-ordered with the objective that different images of the same person obtain a low rank.

Late fusion: inverse score position algorithm (ISPA)

ISPA is motivated by IRPA in that the rank is replaced by a distance. In this fusion method, the per-view matching is applied by directly matching the same views between probe and gallery to

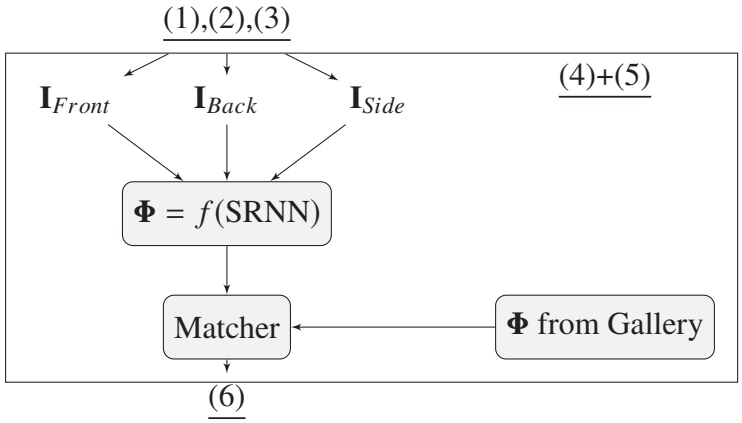


Figure 4.11: Deep Fusion: steps (4)+(5) of Figure 4.4 are shown in detail. The network learns from previously seen examples but it is unclear what it learns.

obtain a view specific distance. By taking a distance instead of a rank, more detailed information on person similarity is utilized. The distance between persons is shorter or longer depending on the similarity, whereas by considering absolute ranks, the distance is disregarded. The mathematics behind IRPA and ISPA is the *harmonic mean*.

Late fusion: product rule (PR)

Furthermore, we employ the product rule (PR). Previous works on biometric multi-modality fusion [Kittler et al., 1998, Alkoot & Kittler, 1999] have demonstrated that the PR adapts well to input data of various scales and does not require extensive normalization of the data [Zheng et al., 2015b]. The resulting distance between a probe and one gallery sample after fusion can be calculated as follows:

$$D(\mathcal{P}, \mathcal{G}_i) = D(\mathbf{\Phi}(I_{\mathcal{P}_{Front}}), \mathbf{\Phi}(I_{\mathcal{G}_{i_{Front}}})) \cdot D(\mathbf{\Phi}(I_{\mathcal{P}_{Back}}), \mathbf{\Phi}(I_{\mathcal{G}_{i_{Back}}})) \cdot D(\mathbf{\Phi}(I_{\mathcal{P}_{Side}}), \mathbf{\Phi}(I_{\mathcal{G}_{i_{Side}}})), \quad (4.7)$$

where $\mathcal{P}$ is the queried probe person, Φ is the image representation in latent space, $\mathcal{G}_i$ a currently compared person from the gallery and D the corresponding Euclidean distance.

Also in this fusion method (cf. Fig 4.10), the per-view matching is applied by directly matching same views between probe and gallery to obtain a view specific score. Results of view specific scores are then fused with the product rule.

Deep fusion: recurrent neural network (SRNN)

In addition to rule-based fusion, we exploit learning-based fusion via a recurrent neural network, where the fusion scheme is learned from training data, see also Section 3.2.2 and Figure 4.11. Instead of multiple consecutive samples used in the original fusion method [McLaughlin et al., 2016], multi-view samples, as in the aforementioned fusion methods, are applied. View

(a) 3m camera height (b) 5m camera height (c) 8m camera height

Figure 4.12: Fisheye images recorded with different camera heights. Persons on the ground plane are in $\mathcal{R}$ oriented quiet similarly to a watch hand.

information is encoded as prior knowledge in the network, and higher learning efficiency is expected. Note that in our setup, optical flow, which is modeled in the fusion method and contributes to multi-shot PRID to use temporal context, is no longer beneficial due to the lack of matching flows.

In this fusion method, per-view matching is not applied, since the SRNN learns from data how to build a feature vector. The pre-defined input-view-order helps the network to focus on how to fuse intermediate feature representations to an embedding. More specifically, the RNN-output-vectors from a front view, a back view, and a side view for one person are averaged in the *temporal pooling* stage (cf. Section 3.2.2). Thus, the processing can be interpreted in away that a *"per average-view matching"* is applied, and a fusion of views was done by the RNN to obtain an average-view.

4.7 Discussion of the approach

In this section, we discuss our approach from a general level, including the restrictions of our approach.

Projection alignment: One central pillar of the proposed approach is the projection alignment step (cf. Sec. 4.3). However, even though projection alignment is applied, resulting images will inherently show decreasing quality with increasing distance to the principal point of the fisheye image. Mapping from a few pixels of squeezed fisheye image space to a large area in the central projection image space of the virtual camera will inherently result in some pixel-intensity-interpolation along with under-sampling of the signal, and therefore increasing ambiguities. This means the quality of person-appearance (i.e. detailed person information) decreases from the shoes to the head, see Figure 4.12.

Data-driven vs. modular processing: In several fields such as computer vision, robotics, and artificial intelligence, most recently proposed methods are based on end-to-end learning and outperform handcrafted modular pipelines by a noticeable margin. With respect to our presented modular processing pipeline (cf. Fig. 4.4), which does not contain end-to-end learning, a design decision was taken, which is discussed in the following section:

Modular pipeline: The benefit is that well known state-of-the-art components can be used which proved to be efficient in the past. For our proposed pipeline (cf. Fig. 4.4), the feature extraction method is an example. Furthermore, a modular pipeline allows the integration of prior knowledge directly into the pipeline, which is provided, e.g. based on handcrafted auxiliary strategies such as the projection alignment of our approach. The disadvantage of a modular pipeline in comparison to end-to-end learning is the missing evidence that this is a high-performing representation; instead, it is an anthropogenic representation. With respect to our proposed approach, there are i) the three view categorization which is similar to the human anatomy, ii) the sampling strategy which selects images of high confidence, and iii) the arrangement of the processing stages.

End-to-end learning: The benefit of end-to-end learning is that a network can optimize itself to obtain a high-performance representation from training data which is not necessarily correlated with any human intuition. The major disadvantage of an end-to-end learning method in comparison to a modular pipeline is the huge training dataset, needed to obtain a high correlation between data used for training and data available during inference; with little training data, this method will inherently not perform well. This means, among other issues, a network has to learn from data how to orient persons for perfect matching and how to handle distortions. However, the primary issue is that this is learned from data which can lead to a bias for general use cases when a correlation between training and testing is decreased, e.g. if the camera pose applied varies from the pose used for training, or the camera lens or scene illumination [Bak et al., 2018] are changed. These facts are underlined by applying state-of-the-art networks, trained on one huge dataset, and evaluating the performance on another huge dataset where poor performance is found.

There is no question that end-to-end learning performs better if a high correlation is given between training and test data and a lot of training data is available. But it remains unclear how to obtain these large amounts of training data for applications in the open world. From a practical point of view, one huge issue is creating a sufficiently large and fully annotated dataset for end-to-end learning, which shows a wide variety of samples to let a network generalize on datasets of different camera poses, and in particular, for different backgrounds.

Since we do not have a large fisheye-image-dataset, but rather solid prior knowledge, we decided to design a modular pipeline. To test whether this is a valid design decision, we compare our approach to an end-to-end learning approach in Section 5.5.6.

Limits of the approach: Finally, we discuss two important limits of our person re-identification approach. Obviously, our presented approach is not able to differentiate between twins showing the same appearance. However, this limitation is valid for most appearance-based person re-identification approaches. Furthermore, face recognition can also be fooled by twins.Besides, persons i) have to pass by using our approach approximately the half field of view in fisheye camera images, to be available in a front, a back and a side view, or ii) need to pass a path that allows for multiple views. It is much easier to obtain multiple views of persons than by employing central projection cameras. However, multi-view can not be guaranteed for arbitrary scenarios. For badly selected camera locations on the ceiling in a scenario, fewer views are available, and our approach can then reduce to a multi-shot approach.

Chapter 5

Experimental evaluation

In this chapter, an extensive experimental evaluation of the proposed approach is conducted. Respective sections present the individual goals, the used datasets, conducted experiments and results, and a conclusion including a discussion.

5.1 General structure of this chapter

In Section 5.2, the benefit of a multi-view strategy compared to strategies such as single-shot or multi-shot is evaluated. The focus is on a low camera pose (front, back, and side views of persons) as available after the projection alignment stage of the proposed pipeline. Furthermore, different feature extraction methods and fusion methods are evaluated to find the best one which is subsequently used to evaluate the proposed fisheye PRID approach in Section 5.5.

In Section 5.3 person re-identification with purely bird's eye views is evaluated, to confirm the design decision to omit bird's eye views in our PRID pipeline.

In Section 5.4, different experiments are conducted to study the influence of the data on PRID. First, data sampling during training is evaluated. Second, experiments are conducted to investigate image quality. In particular image compression and image blurring are studied. Image compression is important to keep image transmission costs low and is typically applied by real security cameras, whereas image blurring occurs with increasing distance from persons' shoes to the head in the images available by using our hardware setup.

Whereas Section 5.2 to Section 5.4 evaluate some basic assumptions of the proposed approach and design decisions, in Section 5.5 the performance of the complete approach for a monocular fisheye camera network is evaluated.

5.2 Multi-view investigations

In this section, we evaluate the proposed multi-view fusion methods and demonstrate the advantages of our multi-view strategy compared to the classical single- and multi-shot strategies for person re-identification. Finally, the best feature extraction method and fusion method are selected for the experiments in the following sections. For a general comparison, we use standard feature extraction methods, which cover both handcrafted and deep features (cf. Sec. 4.5). The feature descriptor and metric learning method model are then fixed to compare single-shot, multi-shot, and multi-view strategies. Thus, steps 4 to 6 of our pipeline (cf. Fig. 4.4) are evaluated in this section. Note that rather than comparing the baseline strategies we focus on the potential of using additional views employing the different fusion schemes.

As handcrafted feature extraction method, the *Hierarchical Gaussian Descriptor* (GOG, cf. Sec. 3.2.1) [Matsukawa et al., 2016] is used in combination with XQDA [Liao et al., 2015] for metric learning. Two different deep feature methods are investigated: i) The method of [Hermans et al., 2017] (TriNet, cf. Sec. 3.2.2) uses a ResNet50 [He et al., 2016] as the backbone, and a triplet loss function with hard-negative mining. ii) In contrast to both other methods, *Recurrent Convolutional Network for Video-based Person Re-Identification* (SRNN, cf. Sec. 3.2.2) [McLaughlin et al., 2016] architecture proved to be an efficient multi-shot fusion method by aggregating features of images acquired one after another. This method is used to quantitatively evaluate if the same method outperforms the other methods by using multiple views. Since no image sequences are used, optical flow, which is used as additional input in [McLaughlin et al., 2016], is disabled for our multi-view experiments.

Note that many feature extracting methods exist. To the best of our knowledge, GOG shows the best performance for person re-identification when using handcrafted features. SRNN, has proven to be useful for aggregating images over time, and data is used to learn the fusion. TriNet, shows high performance across different datasets. Other highly sophisticated methods perform better on particular datasets. However, in preliminary experiments we found that TriNet also performs well for datasets with much higher camera elevation and large viewing angle.

5.2.1 Datasets

Most of the available PRID datasets (cf. Appendix A.2), do not provide labels for views of persons. For our experiments we annotated the Market-1501 dataset with this additional view information. Each image is assigned with a label of "front", "back", "side", and "neither". The original Market-1501 dataset is comprised of 1501 unique persons with over 36,000 samples. After annotation, the dataset is re-mapped (cf. Fig. 5.1) to a multi-view dataset, called MuVi-Market, four single-shot datasets SiSo-Market (front, back, side and random), and a multi-shot

Figure 5.1: Illustration for our Market dataset re-mapping. Views: f= front, b=back, s=side. Timestamps: t_1<t_2<t_3

dataset, called MuSo-Market, with random views and images taken one after the other[1]. The images are divided into two groups, one for probe and one as gallery set, respectively. By following this procedure, we obtain 800 unique person IDs that occur in each of the shot- and view-dependent datasets. Due to the number of unique IDs, we call the dataset Market-800.

In addition, we recorded a new dataset with 85 unique person IDs from three cameras mounted at a height of eight meters. Three views (front, back, and side) per person per camera are manually detected and annotated with bounding boxes. Compared to the Market datasets, camera positions are much higher, and changes in appearance are significant (see Figure 5.2). Note, for this dataset persons are guided in the scenario (cf. Fig. 4.3 (b)+(c)) and a central projection camera is used to study the general influence of multi-view re-identification. Further challenges and properties of our new dataset are i) various accessories like scarfs, baggage, and open jackets can be observed, which change the view specific appearance (50% of the persons in the dataset), ii) uniforms and similar clothes are used by different persons (10% of the persons in the dataset), iii) significant illumination changes across cameras due to automatic camera exposure control have been experienced. Similarly, as for the Market-800, so-called MuVi-Intern and SiSo-Intern datasets are generated for single-shot and multi-view experiments but, in contrast to the Market-800, due to hardware restrictions and annotation effort, not for multi-shot.

5.2.2 Training and inference procedure

To recap, in the following experiments different feature extraction methods (GOG, TriNet, SRNN) and fusion methods (early fusion, product rule, ISPA, IRPA, deep fusion) are evaluated to find the performance gain by using a multi-view strategy, and to find the best performing

[1] Note, $\Delta_1 = t_2 - t_1$ and $\Delta_2 = t_3 - t_2$ are unknown, not fixed, and given by the underlying public dataset

(a) Persons (1-85) observed by camera 1

(b) Persons (1-85) observed by camera 2

(c) Persons (1-85) observed by camera 3

Figure 5.2: Our multi-view dataset MuVi-Intern: persons in three views (side, front, and back). For visualization, all person images are downscaled to an equal image resolution and automatic image enhancement is applied for printing.

combination. Whereas for the Market-800 dataset, all three feature extraction methods were applied, only the non-deep-learning based method, i.e. GOG and XQDA, is applied to the internal dataset, due to the very limited size. Furthermore, for the handcrafted method in combination with late fusion, we decided to train one metric per view combination (front-front, back-back, side-side), instead of one metric which is trained on arbitrary view combinations. For *TriNet*, we use the source codes provided by the authors, starting with the ResNet50 [He et al., 2016] trained on ImageNet [Deng et al., 2009]. We did not directly use the model provided by the authors since it potentially was learned with an overlap of Market-800 test data. In contrast to GOG and XQDA,[2] and late fusion, training for TriNet is done without view dedicated learning since using the hard negative and positive mining we expect that the network learns this information itself. Additionally, we can teach the network with many more samples since three person-views are used for the batch hard loss. Note, in contrast to the original publication [Hermans et al., 2017]; we train the network with a significantly smaller training data size, around 13% of the original samples, since our Market-800 dataset consists of only a subset of the original Market-1501 dataset. Consequently, we expect a performance drop, even if data augmentation is used. The original *SRNN* aggregates information over time and uses optical flow for multi-shot samples. For our multi-view experiments, we feed data in a fixed order (front, back, and side view) into the network to perform the fusion. By using the SRNN we expect that the model learns which colors and patterns are available from all three views and which are important to re-identify individuals.

Following the VIPeR [Gray et al., 2007] benchmark protocol, we apply cross-validation on both datasets, the Market-800 and the internal dataset, where 10 different dataset splits, are randomly generated with 50% IDs for training and 50% for testing at each split. Averaged rank#1 over all splits of each dataset are reported, where rank#1 accuracy is the probability of finding the correct match in the first rank of the Cumulative Matching Characteristic (CMC) curve [Gray et al., 2007]. For multi-shot experiments we take the fusion equations from Section 4.6 and replace the imaged views (f,s,b) with images from time steps (t_1,t_2,t_3).[3] Note, we used the publicly available implementation of *SRNN* [McLaughlin et al., 2016], changed the dataset, and fixed the train and test splits to the one used for the other experiments. Besides, the optical flow is disabled for multi-view experiments.

Furthermore, we highlight that feature vectors are extracted for TriNet regardless of a view as prior-knowledge. This means, there is only one feature extractor to obtain an embedding and not a separate one for each view. Two alternatives could be: i) Four networks. Also, one for each view, plus one for matching arbitrary view-combinations. ii) A particular network architecture using different view specific branches [Eberle, 2018]. We motivate our decision to use only one feature extractor by the fact that we have three times as many samples for training, and in particular, can provide much more hard samples to teach the network. Further, some research

[2] Note, for GOG and XQDA we set the hyper parameters exactly as in the the publication of the authors.
[3] Note, $\Delta_1 = t_2 - t_1$ and $\Delta_2 = t_3 - t_2$ are unknown, not fixed, and given by the underlying public dataset

Table 5.1: Rank#1 recognition rate (%) of conducted experiments, results are averaged over ten dataset splits. No multi-shot is performed for the internal dataset due to lack of data. Note, the results of the SRNN are only shown in Fig. 5.3, since a different feature extraction method and a different fusion method is used, both are not related to the method used in this table. (concatenated feature vectors = CFV, inverse rank position algorithm = IRPA, inverse score position algorithm = ISPA, product rule = PR)

Fusion \ Feat. Ext.	GOG+XQDA	TriNet	GOG+XQDA
Single-Shot	SiSo-Market		SiSo-Intern
Random Views	40	74	43
Front Views only	65	78	59
Back Views only	64	78	64
Side Views only	51	73	58
Multi-Shot	MuSo-Market		-
CFV	55	73	-
IRPA	58	74	-
ISPA	62	77	-
PR	61	76	-
Multi-View	MuVi-Market		MuVi-Intern
CFV	83	87	73
IRPA	83	91	80
ISPA	**87**	**95**	**82**
PR	86	94	81

papers [Zheng et al., 2016b] claim that e.g. the VIPeR dataset [Gray et al., 2007], which has $i = 316$ IDs and $n = 632$ images for training and testing respectively, is too small to apply a deep learning based feature extraction method. To overcome such a data issue with our datasets, we employ a feature extractor which takes images without any view attribute. Thus, i persons which are available in three views in two cameras, provide $i \cdot 3 \cdot 2$ images for training.

5.2.3 Evaluation and discussion

Comparison of single-shot, multi-shot and multi-view: In this paragraph, we compare our proposed 3-view set fusion strategies with the conventional single-shot and multi-shot matching strategies. The three feature baselines are combined with the corresponding strategies of the three categories. The result is shown in Table 5.1 and Figure 5.3. Since multi-shot uses more person information than single-shot with a random view, a much higher performance is obtained; for TriNet 77% vs. 74% rank#1, and for GOG and XQDA 62% vs. 40% rank#1. However, using only a single view (front or back) gives for both approaches a better performance than multi-shot (78% vs. 77% rank#1 for TriNet and 65% vs. 62% rank#1 for GOG and XQDA), which indicates that a restriction of views is useful. Compared to the other two categories, our multi-view approaches improve the performance further, independent of the feature baseline.

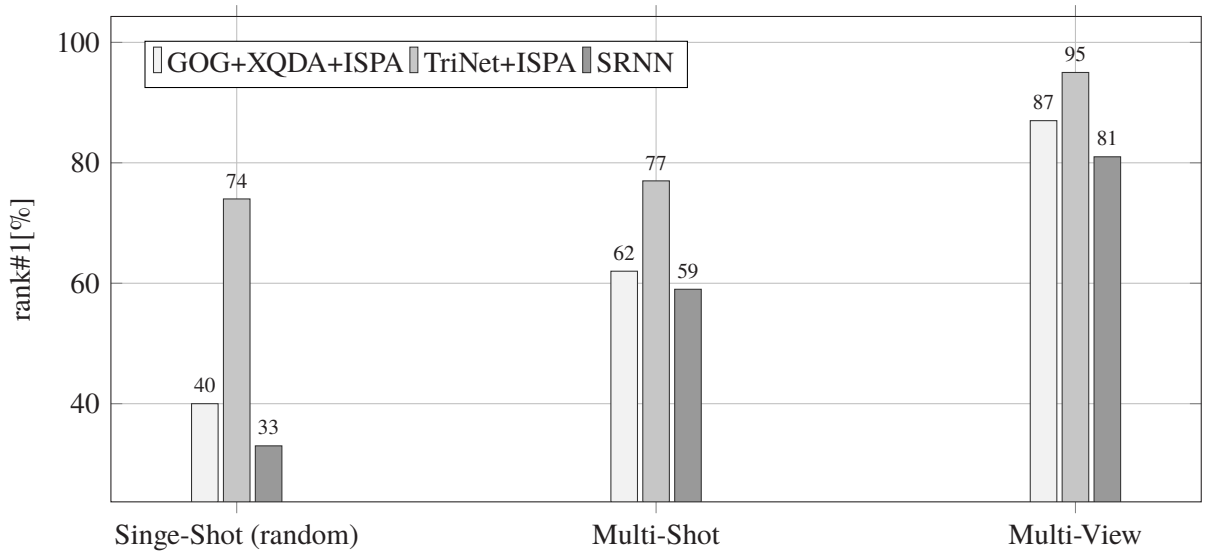


Figure 5.3: Comparison of different rank#1 (%) for the Market-800 dataset. Best performing rule-based fusion and the SRNN fusion are illustrated for single-shot, multi-shot and multi-view.

For example, with GOG and XQDA, multi-view approaches increase the maximum rank#1 by +47% (40% vs. 87%) and +25% (62% vs. 87%) compared to single-shot with random views and multi-shot. Similar trends can be observed for TriNet (+21% and 18%, respectively) and SRNN (+48% and +23%, respectively). These improvements clearly show the impact of additional person view information for the re-identification task. With deep features, the improvement is less significant, probably because they extract richer information at mid-level already, as shown in [Geng et al., 2016]. At least for handcrafted features view knowledge seems to be also very useful for single-shot matching since rank#1 could be improved from 40% random view matching to 65% for front view only matching. The results for the internal dataset show that the performance slightly decreases compared to the same strategies on the larger Market-800 dataset which is probably caused by the small training dataset, combined with the much higher camera position.

Deep features vs. handcrafted features for fusion: In this paragraph, we examine the impacts of multi-shot and multi-view fusion on both deep learning and the handcrafted feature baselines. For GOG and XQDA, multi-shot and multi-view increase the performance significantly compared to single-shot random views. However, for TriNet there is no significant difference between single and multi-shot matching. The SRNN seems to be an effective fusion for the Siamese features since rank#1 was improved by +25% (33% to 59%). The performance variations between feature baselines were reduced from 38% (max. performance minus min. performance for single-shot; 78% - 40%) to 12% (max. performance minus min. performance; 95% -83%) after multi-view fusion. These results show that view information is important, and independent of the type of PRID features learned from images. Therefore, both strategies can be very effectively combined.

Comparison of single-shot and fusion methods: The proposed fusion schemes have been applied to both multi-shot and multi-view experiments. As shown in Table 5.1, ISPA and PR are the most effective methods. SRNN improves results of Siamese features from 33% for single-shot to 59% for multi-shot to 81% for multi-view. Note that we cannot compare these results to other publications directly since we use the re-mapped dataset. By considering the complexity and generalization, ISPA and PR are recommended to be used, as they are simple to perform, and no data-specific training is necessary. It is worth mentioning that the best performance of our fusion on Market-800 is on average 95%, which is at a very high level.

Discussion: The experiments indicate a much higher rank#1 recognition rate for multi-view compared to the other strategies independent of datasets and feature extraction method. By comparing the improvements on the re-mapped Market dataset (+3% from single-shot with TriNet to multi-shot with TriNet [4] and +18% from multi-shot to multi-view [5]), and the internal dataset (+39% from single-shot to multi-view [6]), one question could arise: is the proposed multi-view approach still beneficial if more training data is available for baseline approach? As we saw, multi-view information is important, and irrespective of featuring extraction methods. Consequently, our multi-view strategy can improve the re-identification performance further, even though more training data is available.

In summary, the results of conducted experiments indicate TriNet performs better than GOG along with XQDA or the SRNN, and ISPA gives in combination with TriNet the best rank#1 performance. Thus, the combination of TriNet and ISPA is used in all following multi-view experiments.

5.3 Bird's eye view investigations

In this section, bird's eye views (BEVs) of persons are studied, which are provided by the proposed hardware setup whenever persons cross the central fisheye image region. Considering that the central image region provides the highest resolution, the BEV data allows the observation of fine-grained person heads. These glimpses were not investigated in related work on PRID and could be beneficial for the re-identification. Therefore, we i) experimentally evaluate whether BEVs are useful for our task, ii) give a statistic of BEVs in fisheye images, and provide a comprehensive discussion with respect to a practical application.

[4] 74% vs. 77% rank#1 on Market
[5] 77% vs. 95% rank#1 on Market
[6] 43% vs. 82% rank#1 on the intern dataset

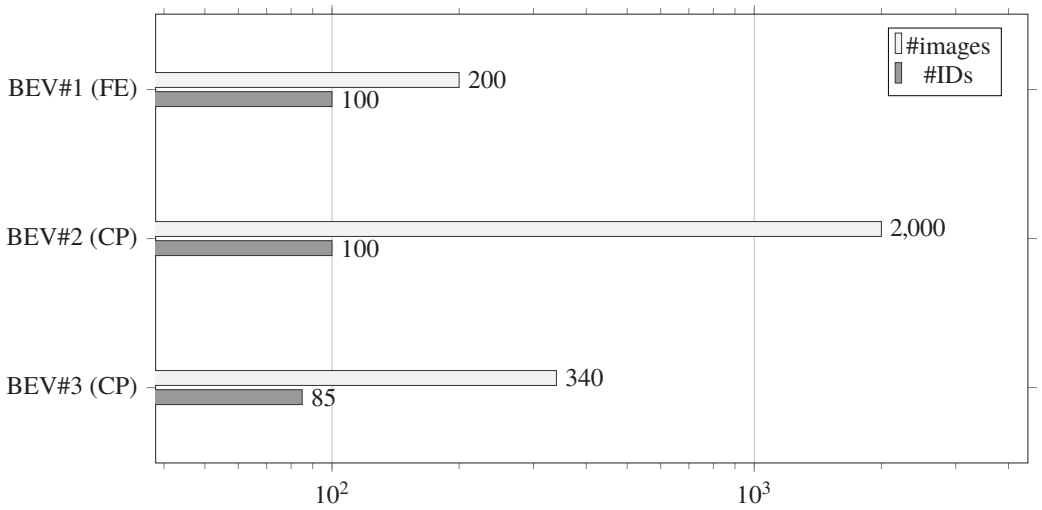


Figure 5.4: BEV dataset overview: the absolute number of ID and images is illustrated. Peculiarities are highlighted: fisheye (FE), and central projection (CP).

5.3.1 Datasets

To the best of our knowledge, only one dataset (TVPR [Liciotti et al., 2017]) for bird's eye view person re-identification (BEV-PRID) exists. TVPR which consists of 100 persons and allows the extraction of anthropometric features[7] for PRID.

Due to the different focus of [Liciotti et al., 2017], there is an overlap between persons used for training and testing which is important for evaluating appearance-based person re-identification. Further, only one camera is used for one scenario for one camera pose. Since TVPR is, thus, not applicable for our goal, we created four new datasets to study BEVs for appearance-based person re-identification (cf. Figure 5.4 and Figure 5.5):

① BEV#1 is recorded by two fisheye cameras and consists of 100 persons and 200 images. Each person was directed to pass the two central image regions and is therefore recorded in two BEVs. The detections of persons were obtained by applying a Mask R-CNN [He et al., 2017] based on the implementation and training of [Girshick et al., 2018].

We used a camera height of 4m, thus, a person appears in a tiny image region containing the bird's eye views. However, applying a smaller camera height reduces the ability to observe front, back, and side views of persons.

Furthermore, persons are only visible in a BEV for a very short time, resulting in only a few image frames. Moreover, due to the movement of persons affecting a huge number of pixels in BEV images, short exposure time and high frame rate are mandatory to reduce motion blur. Higher frame rate is critical in terms of computational costs and practical application. Note, for BEV#1 we only use intra-coded frames (I-frames from

[7] anthropometric features are body measures of persons

Figure 5.5: Excerpt of BEV datasets. Three different persons are randomly selected and horizontally depicted per dataset. (BEV#1 first row, BEV#2 second row, BEV#3 third row)

h264 video-stream), to slightly improve the dataset quality, by removing the inter-coded frames.

② BEV#2 is taken by two central projection security cameras in nadir direction and consists of 100 persons and 2000 images. Even though, central projection cameras are employed, the experiments allow the conclusion for the central image region of a security fisheye camera, where distortions can be neglected, but images are compressed to keep transmission costs and storage space low. Annotations for this dataset were carried out manually. Furthermore, many more samples are available as the whole field of view can be used to obtain BEVs compared to BEV#1.

③ BEV#3 is acquired by two central projection consumer cameras and consists of 85 persons and 340 images. The camera height was larger, approximately six meters instead of approximately four meters as for BEV#2. In contrast to the security camera images of BEV#2 the images look sharp. Similar to BEV#2, the experiments allow the conclusion for the central image region of a fisheye camera. The detections of persons were obtained by manual annotations. Furthermore, in comparison to BEV#2 different persons are recorded and non-homogeneous backgrounds are used.

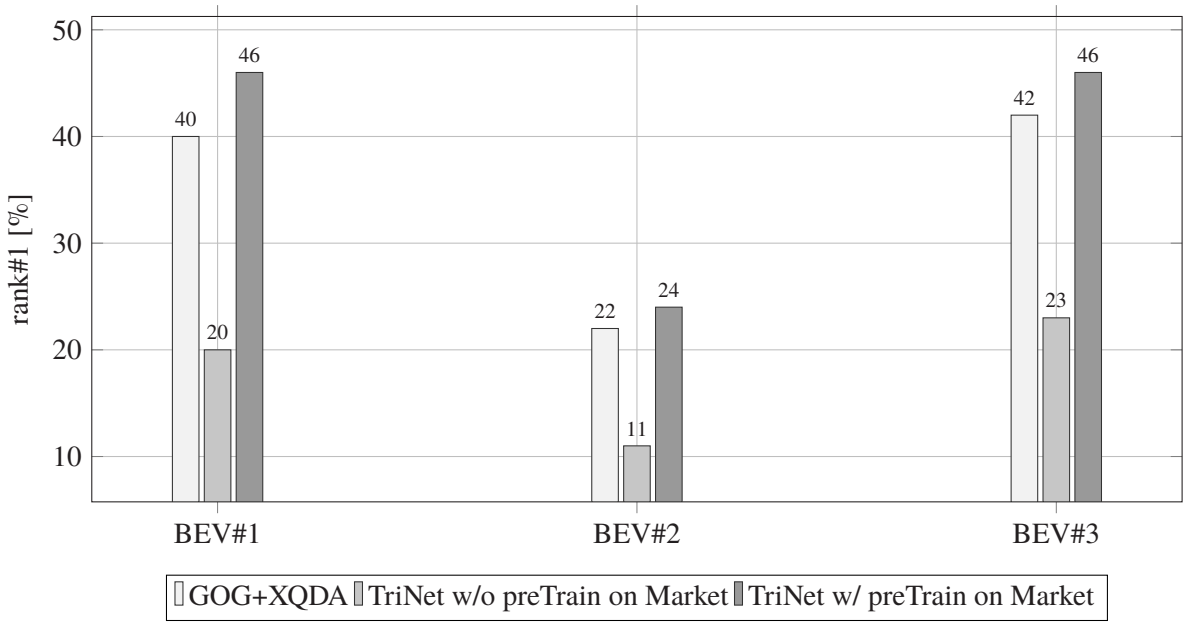


Figure 5.6: Comparison of different rank#1 (%) for the BEV experiments

5.3.2 Training and inference procedure

GOG and XQDA as handcrafted features with metric learning, and *TriNet* as deep features are used with the same parameters as in the experiments of Section 5.2. 50% of samples are used for training, 50% are used for testing, and ten splits were applied for cross-validation. Due to the aspect ratio of BEV images, all images are sampled to 256x256 pixels before they were used as input. All *TriNet* experiments are conducted on a model, which was initialized with weights from a training on the ImageNet dataset and a further training on Market-1501, before the network was trained for BEV data.

5.3.3 Evaluation and discussion

In this paragraph, the results of conducted BEV-PRID experiments are presented, see Figure 5.6. The performance for the real-world BEV data is low compared to the performance obtained on classical re-id data where a performance of more than 80% rank#1 by using 750 test persons is feasible. Further, it is shown that a pre-training of TriNet on Market-1501[8] also helps for BEV data. BEV#2 shows the lowest performance which is likely caused by more motion blur in the dataset. BEV#1 and BEV#3 show approximately the same performance. A little less than every second person was correctly re-identified. One further effect is that the performance gap between the handcrafted feature (GOG) and the deep learning-based method is hardly visible compared to the results in Table 5.1.

[8] Note, this is done additionally to a pre-training on ImageNet.

Figure 5.7: Exemplary matching result from the BEV experiment for BEV#2. The image on the left (black border) shows the probe, whereas the five following images show the most probable matches ordered by decreasing similarity. Match one to three are false matches (red border), match 4-5 show the correct match (green border). The person in the correct match is shown in a different pose and therefore complicates the matching.

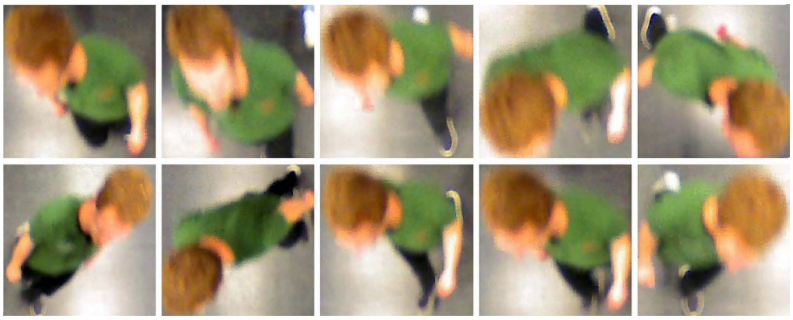


Figure 5.8: Illustration of different variations in which persons typically appear in BEV images. Holistic feature extraction methods are likely to fail, whereas person-pose-based methods could help. The images of the person are from the BEV#2 dataset which has unsharp and blurred properties.

Besides, we assume the poor performance is in particular caused by much more appearance variation caused by varied orientations of persons. Figure 5.7 depicts one matching example and Figure 5.8 one person in different BEV images. In Figure 5.7, a wrong match seems more similar than the correct match. Taking into account that hair and skin color are entirely different, it seems like the same color distribution for the same parts in an image is more decisive than distinctive colors of body parts. Therefore, person-pose integration for BEVs is much more critical than for lower camera poses where persons are horizontally projected into images.

To conclude, high-resolution bird's eye views are provided by the proposed hardware setup whenever persons pass the central field of view, which is not easy to achieve by using arbitrary camera locations on the ceiling. Even though persons are directed to pass the central field of view as in this dataset, the results of conducted experiments show a low re-identification performance on datasets of a trivial difficulty level ($n = 50$ persons for testing). As shown, in BEVs persons are available in many more variations compared to the classical re-identification datasets (cf. the head and body positions, and orientations in Figure 5.8). We believe these different variations may benefit from a pose-based modeling approach to integrate the person pose as prior knowledge into the matching. However, commonly used public person-pose-keypoint-datasets, such as *MPII Human Pose Dataset* [Andriluka et al., 2014] or the *COCO Keypoint Detection Task* [Lin et al., 2014], consist of horizontal projections of persons into the images. Therefore, they are not an option for deep learning based keypoint-detection in BEV images, and in particular, they are not an option to train a proper person-pose-detection-branch of a network for BEV-PRID. Thus, a very important fact is that a large amount of nadir view

images are not publicly available for training the person re-identification and the person pose detection task, jointly, today.

Despite the fact that BEV images are available in the recorded scenario by using our proposed hardware setup, we do not follow the BEV-PRID further due to the issues mentioned above.

5.4 Influence of data

Enough training data and a sufficient image quality are mandatory prerequisites in applying learning-based approaches. Whereas in the two previous sections, data is taken "as it is" and randomly divided into train and test splits, in this section, the influence of training samples is evaluated. In particular, sample selection and the effect of image quality are studied regarding TriNet, which showed superior performance in the experiments of Section 5.2. The following experiments are conducted:

Data sampling is studied to answer the question if the final re-id performance for a given feature extraction method is only a question of the amount of training data, or rather a question of the number of persons within the training dataset, of images per person in the dataset, or of the number of images regardless of other points.

Image quality is studied to evaluate the influence of image details. We decided to investigate two aspects i) the influence of *jpg compression* and ii) the influence of *image blurring*:

i) jpg compression is investigated to understand the influence of image details on PRID. For public PRID datasets, the recording setup is mostly unknown as well as potential post-processing steps. However, since our new fisheye PRID datasets are recorded with *real security cameras* and many public datasets look as if recorded with consumer cameras, the influence of image compression is important to understand re-identification performance differences. Real security cameras typically use jpg compression to produce bitrate optimized videos and images to keep transmission costs and storage space low. Therefore, the image quality is degraded.

ii) image blurring is studied for PRID in de-warped fisheye images to better understand the effect of increasing blurring with increasing distance from the principal point (cf. "Projection alignment" in Section 4.7).

5.4.1 Datasets

In this section, all experiments are based on the Market-1501 dataset [Zheng et al., 2015a], which is discussed in detail in Appendix A.2. Since in this section no multi-view dataset is needed, we use the public dataset instead of our datasets to obtain as many images as possible for training and testing. Consequently, we allow for comparability with related work.

5.4.2 Training and inference procedure

To investigate the influence of **data sampling** and **jpg compression**, the following experiment is conducted based on the TriNet. Different strategies are used to train TriNet with different subsets of the Market-1501 training dataset.

1. One set consists of the complete Market-1501 training dataset and gives the best performance for the TriNet, since it consists of all training images from the Market domain. It allows the determination of the upper-performance border.

2. Ten further sets are sampled with 10% to 100% person IDs in 10% increments of the original training dataset. These sets are used to determine the influence of using more person IDs for training. The number of images per individual is given by the underlying dataset.

3. The next ten sets are created by keeping the percentage settings of (2), to evaluate the influence of using more images. Thus, more and more images are used for training but there is no linear relationship between more images and more persons. Rather, a random choice is applied.

4. In addition, ten further sets are sampled, by keeping the percentage settings of (2), which means that with an increasing number of images more and more persons are sampled. A linear relationship between the number of images and the number of persons is given in this case.

5. For the second goal, the influence of **jpg compression**, ten models are trained, with 10% to 100% jpg quality level[9] in 10% increments, to analyze the performance drop by using low-quality training and test data. Note, jpg divides images into 8×8 blocks and applies YCbCr colorspace transformation, a discrete cosine transformation, and quantizes the resulting coefficients. The lower the quality level, the wider the quantization step size, and the less detailed information remains. More details about jpg can be found in ISO/IEC 10918-1. In these experiments, for each quality level, all images of the Market-1501 dataset are converted with libjpg[9], which is a standard software library for jpg, for training and testing to the same respective quality level. Thus, the performance difference between different quality levels, for one and the same test and train split compared to the original dataset, can be evaluated by conducting the described experiment.

To study **image blurring**, the following experiment is conducted: person patches from the Market-1501 dataset with dimension 128×64 pixels are divided into four horizontal regions with dimensions $(4 \times 32) \times 64$ pixels. Blurring using a box filter (averaging the pixel values per channel) with pre-selected kernel size (κ), details are given later, is applied to the regions to

[9] `http://refspecs.linuxbase.org/LSB_3.1.0/LSB-Desktop-generic/LSB-Desktop-generic/libjpeg.jpeg.set.quality.1.html`

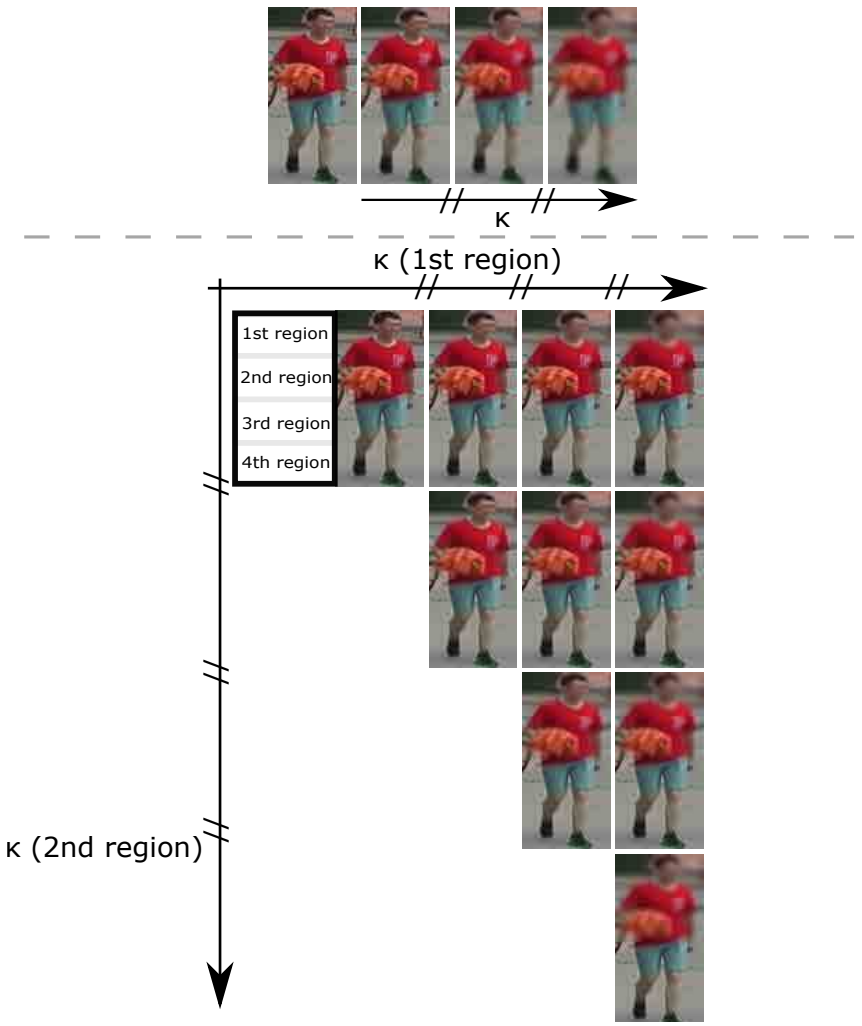


Figure 5.9: Illustration of blurring to study the influence of decreasing quality from shoes to head. Top subfigure shows one an the same image blurred with $\kappa = [3, 5, 9]$. Bottom subfigure shows one image where $\kappa = [3, 5, 9]$ is applied on first and second region independently.

simulate the blurring effect of interest in a simplified manner. For each blurring combination, one model is trained with TriNet and evaluated, i.e. all images are processed to the same blurring combination for training and testing, respectively. This allows the investigation of the blurring influence whereas the dataset is always the same and the original dataset without blurring can be used as a reference.

To illustrate the experiment, Figure 5.9 shows different blurred images on the top, and different blurring combinations in the first region and the second region on the bottom.

In the experiments, the effect on i) the top region only, ii) the first and second region, iii) the first to the third region, are evaluated with different kernel sizes for blurring. Even though this simulation is only a rough approximation of the real "fisheye-blurring-effect" and does not model an adequate replacement including smooth blurring transition, we believe it allows the study of the impact of blurred image regions.

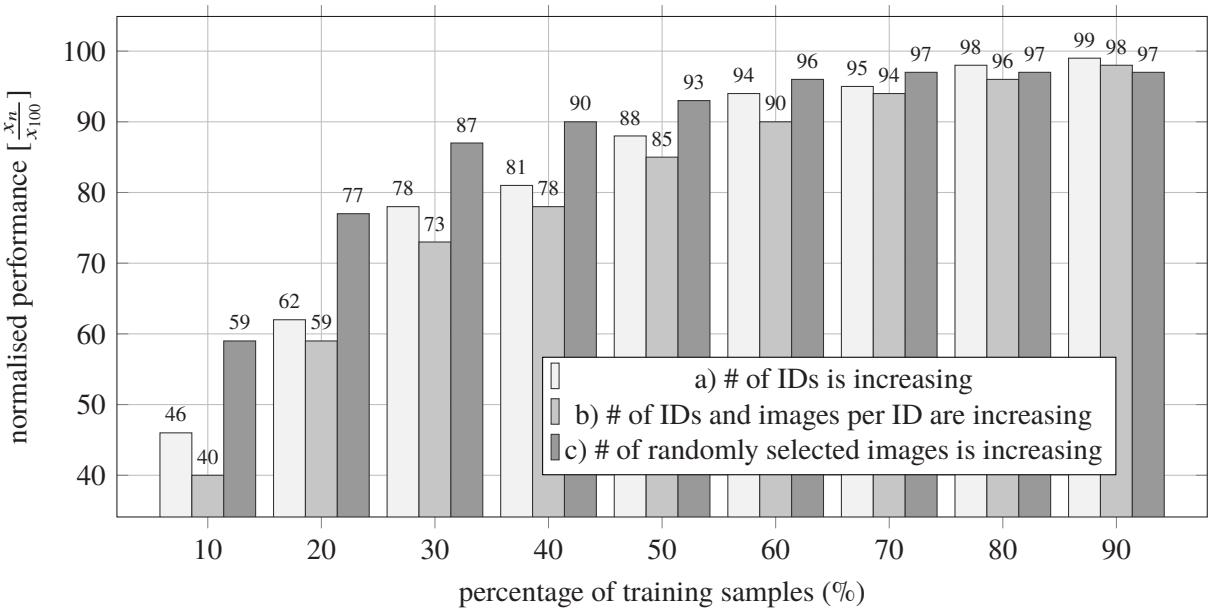


Figure 5.10: Influence of training data size. a) # of IDs is increasing; images per ID given by underlying dataset, see 2), b) # of IDs and images per ID are increasing, no linear relationship, see 4). c) # of randomly selected images is increasing, see 3).

5.4.3 Evaluation and discussion

Data sampling: Figure 5.10 summarizes the results of the conducted experiments to study data sampling. The performance is normalized by the rank#1 obtained by training the network on the full training dataset to directly consider the performance drop between different experiments compared to the maximum performance by using the underlying dataset. Note, the maximum performance of TriNet is 84.9% rank#1 on this dataset.

Using more person IDs is more important for a high person re-identification performance than using more images of the same persons which can be explained by the applied loss function. Since a hard-negative mining strategy is applied, only hard samples within a batch are used to update the network's weights. Furthermore, the figure shows a saturation curve, i.e. the performance seems to be plateauing, which indicates that more data samples will only slightly improve the performance. With increasing performance, more data is needed to improve the network further.

With 60% of images, more than 90% of the maximum performance is reached. Using only 50% of persons, results in 88% of the maximum performance. Note that in contrast to the results related to *deep learning*, experiments from other fields show that larger and more diverse datasets noticeably improve the performance. Thus, we believe that in our case the limiting element is not the architecture, but rather the amount of training data and the variety of the dataset. Two reasons are assumed to be responsible for the depicted early saturation of our experiments: i) A lot more hard samples are mandatory to boost the performance further. Due to the hard-negative mining, only hard samples are important and not the number of person images per ID, which

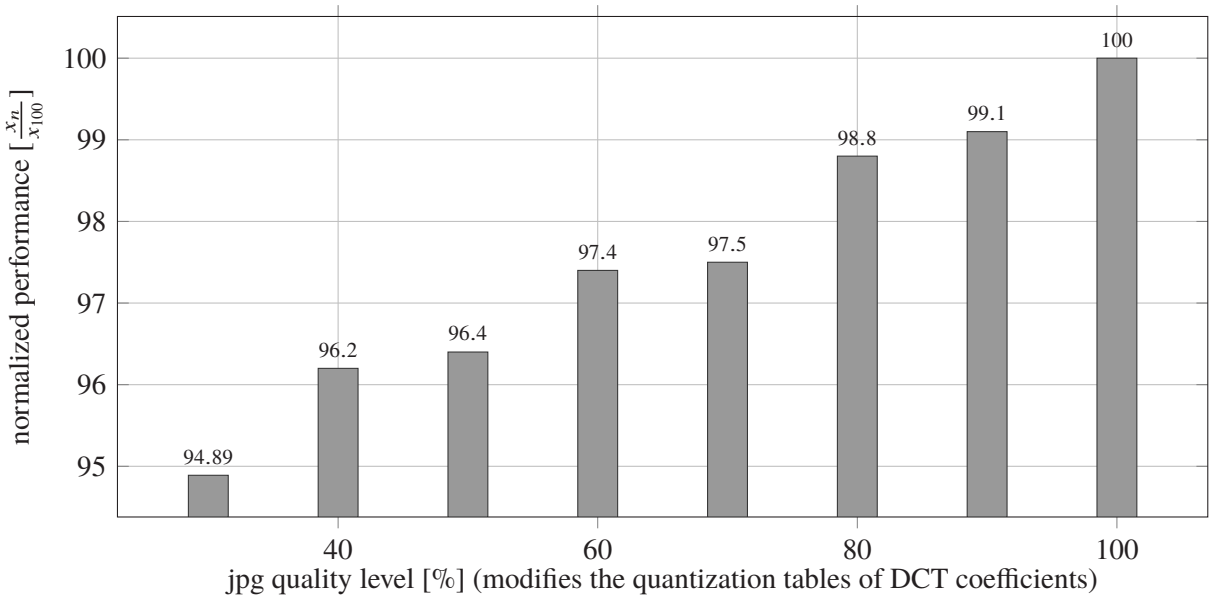


Figure 5.11: Influence of jpg quality (see ISO/IEC 10918-1)

are too simple to optimize the network for better performance. ii) Due to the single-shots and matching of arbitrary view combinations, more variety in person observations and view-invariant person-description are mandatory for person comparison, to describe a person comprehensively for matching. In other words, more views are needed to describe a particular person with one feature vector.

The conclusion for the proposed pipeline can be stated as follows: experiments in the previous section were conducted on a re-mapped and reduced version of Market-1501. Since around 50% of the persons in the complete dataset are used in the re-mapped version (cf. Sec. 5.2) and in particular due to the hard samples (different views), the feature extraction method is able to give around 90% of the maximum achievable performance for the re-mapped Market dataset of Section 5.2. Therefore, the results of experiments on the small dataset are indicative for using fewer samples for training as used in Section 5.2. Moreover, the experiments show that the more persons available for training, the better the re-identification performance.

Jpg compression: In Figure 5.11 the results of image compression are presented. The performance is normalized by the performance obtained by 100% quality level. The influence of the compression has a minor but noticeable effect on the accuracy. However, the curve follows an approximate linear relationship. The quality deterioration from 100% to 30% shows a relative drop of around five percent. This indicates - for this particular dataset - that jpg quality of imaged persons is not as vital as one could assume.

The conclusion for the proposed pipeline can be stated as follows: Even though a reduced quality level is applied due to the employment of a security camera, no significant performance drops were found to be caused by poorer image quality in terms of compression.

Table 5.2: Normalized re-identification performance for experiments with different levels of blurring. Same kernel size is applied to multiple regions to conclude for general application. The rank#1 performance is normalized by the performance obtained by disabling the blurring.

norm. perf. [%]	**blurred regions**		
κ	**1**	**1-2**	**1-3**
3	100	100	100
5	99	98	98
7	98	96	95
9	96	92	89

Table 5.3: Normalized re-identification performance for experiments with different levels of blurring. A different kernel size for the first and second regions is applied to account for de-warped fisheye images. The rank#1 performance is normalized by the performance obtained by disabling the blurring. The table shows a triangle, since region-1 is at least as blurred as region-2.

norm. perf. [%]	κ (1^{st} **region**)			
κ (2^{nd} **region**)	**3**	**5**	**7**	**9**
3	100	99	99	99
5		98	97	97
7			96	96
9				92

We will continue the discussion of this point in Section 5.5. Then, we discuss reasons of low performance on our own real security camera dataset in comparison to the performance on a public dataset.

Image blurring: Table 5.2 and Table 5.3 depict the result for increasing image blurring. The rank#1 performance is again normalized by the performance obtained by non-blurred images.

There are several important findings: i) As long as κ is smaller than five, the performance drop is less than around two percent, compared to the performance obtained by non-blurred images. Since we expect slightly blurred images due to the camera setup and projection alignment step, a κ of five shows qualitatively more blurring (cf. Fig. 5.9) than we expect to obtain by using a fisheye camera and apply projection alignment in $\mathcal{R}$ (cf. Fig. 4.6), this will be underlined later in iv). Whereas in Section 5.2 it was shown, that multi-view allows the improvement of the performance compared to other strategies by a two-digit margin (cf. Table 5.1 and Fig. 5.3), the blurring experiments show that if the image quality is slightly deteriorated, this effect has a minor contribution to the overall re-identification performance of the feature extraction method. Additionally, the results indicate (cf. Fig. 5.11), that detailed image information does not seem as vital for the deep learning-based feature extraction method as could be assumed.

ii) Blurring the top region only, which includes the person's head, has a minor contribution to the re-identification performance (-4%), even though a κ of 9 is selected. This indicates for the selected feature extraction method that the head-region is not very important to improve the

re-identification performance. It is an interesting finding to our approach since the head shows the most blurring in the de-warped fisheye images. However, we claim that the head and in particular, the hair color, and accessories are important to differentiate between persons. Thus, person re-identification can be improved by additionally using an approach that is tailored to handle these features (hair color, hair cut, noserings, glasses, etc.).

Investigating the dimensions of the feature maps from the applied backbone[10], the effect of discarding detailed information is not surprising due to the small feature map resolution. Detailed spatial-information from the high resolution input images[11] is removed by the convolution structure of the CNN.

iii) Blurring the region 1 to 2 drops the performance for a κ of 9 up to 8%. Blurring region 1 to 3 decreases the performance for a κ of 9 by 11%. This indicates that regions 2 to 3 with a κ of seven are more critical for the re-identification than a person's head. Thus, the appearance of clothes has much more contribution to the overall person re-identificaion performance than detailed head information like hair or a hat.

iv) The real fisheye data which we present in Section 5.5, is qualitatively comparable to a $\kappa = 5$ for the first region and $\kappa = 3$ for the second region (cf. Fig. 5.13)[12]. Analyzing the performance for this blurring combination leads to a relative performance drop of less than one percent relative to the performance obtained without blurring. This indicates that decreasing person-quality from shoes to the head is negligible for person re-identification.

In summary, the results of the conducted experiments indicate that, by using a fisheye camera and the projection alignment step, a performance drop is hardly visible as long as images from $\mathcal{R}$ (cf. Fig. 4.6) are used.

5.5 Fisheye investigations

After the evaluation of the fundamental assumptions, design choices, and an investigation of expected weaknesses, the complete pipeline is evaluated in this section. First, an overview is given of the datasets used to investigate our PRID approach. Subsequently, projection alignment, view classification, and the re-identification are evaluated.

5.5.1 Datasets

Most of the available PRID datasets (see Appendix A.2) are acquired by central projection cameras and do not provide labels for person views. Therefore, to compare our multi-view fisheye

[10] ResNet50, stage 1: 62×32pixels; stage 2: 32×16pixels; stage 3: 16×8pixels; and stage 4: 8×4pixels.
[11] 256×128pixels
[12] This finding is based on an inter-rater test [Saal et al., 1980] consisting of 9 persons.

data experiments with large scale multi-view public data[13], we pre-processed the Market-1501 dataset, and, in addition, acquired a new fisheye dataset for our experiments:

i) Market Multi-View (MMV): the publicly available Market-1203 dataset [Ma et al., 2016][14], which is a re-mapped version of Market-1501 [Zheng et al., 2015a] with images from two central projection cameras (one providing the probe images, the other one the gallery images) and manual annotations, was again re-mapped to a multi-view dataset. For training 5413 images showing 934 individuals are used, consisting of various view combinations, but without persons which appear in the three different views. This is a design decision which we took to increase the size of training data and the size of test data. However, for testing dense 3-view sets consisting of all three views per camera are a mandatory prerequisite for a fair comparison with a fisheye camera dataset (see below). In our case, 1614 images of the respective two cameras depicting 269 individuals are used for testing.

ii) Fisheye PRID (FEP) is a new multi-view person re-identification dataset, which consists of 300 persons captured in a camera network of five fisheye cameras. Cameras are mounted at the height of 4m to 5m, the optical axis points in nadir direction. Three cameras focus an open space airport scenario with check-in area and security-check comprising of 85 persons. Two cameras record a retail shop, here 215 persons are available. FEP consists of 5900 images and is divided into 200 persons for training and 100 persons for testing. In the FE images, persons were detected manually and annotated with bounding boxes. As mentioned before, automatic detection was considered to be beyond the scope of this work. Due to the wider field of view of FE cameras and the primary open space in the airport scenario, complete 3-view sets from all 85 people are available. In contrast to the open space scenario, the retail scenario contains a number of occlusions caused by, e.g. shelves. As a consequence, only about 80% of the persons appear in 3-view sets consisting of at least one front, one side and one back view; around 6% appear in one side view with either front or back view but not both (front and back view), and 14% of persons appear in at least one front and one back view. Note, these values are caused by the particular scenario, camera pose, and the occlusions which are typical for a store.

For the recordings the Bosch Flexidome IP panoramic 7000 MP [15] security fisheye camera is used which has the following characteristics:

Sensor: 12MP (1/2,3-Zoll-CMOS); used pixels: 2640 x 2640 (7 MP); lens: 1.6 mm fixed-focus fisheye lens (IR corrected), F2.8 (iris fixed); video compression mode: H.264 video coding, with image quality optimized settings (see [15]).

[13] Note that after projection alignment fisheye images look as if they were taken by central projection. It is only visible that the camera height of the virtual camera is higher compared to public datasets.
[14] https://github.com/charliememory/Market1203-Reid-Dataset
[15] https://de.boschsecurity.com/de/produkte_1/videosystems_1/ipcameras_1/panoramiccameras/flexidomeippanoramic7000m_1/flexidomeippanoramic7000m_1_18936

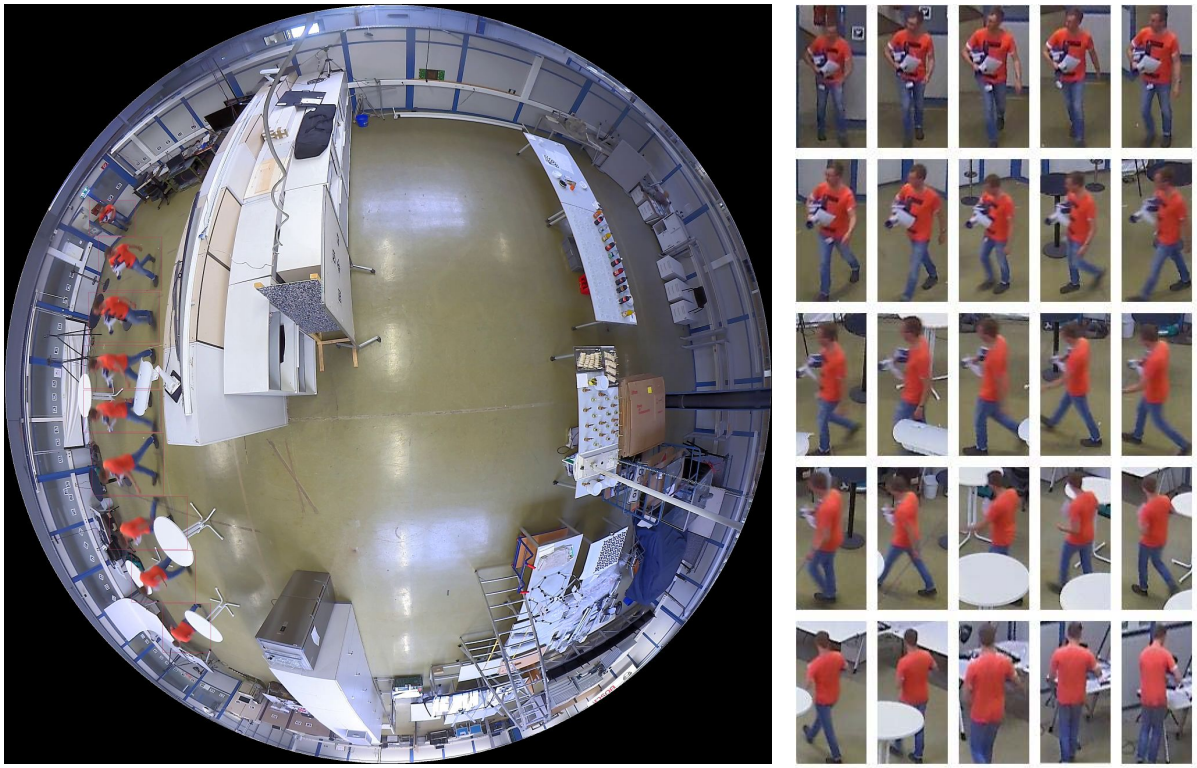


Figure 5.12: Exemplary result of person projection alignment for one observed person. Left: multi-exposure image (video synopsis) of one person in an FE image. Right: multiple person detections manually annotated and then automatically de-warped and aligned. Note that depending on the selected frame rate multiple images of a person are available and only nine are exemplary shown in the multi-exposure image to avoid person overlaps in the image space. On the right 25 aligned images are depicted.

5.5.2 Training procedure

In Section 5.2 it was shown that TriNet shows the best person re-identification performance in combination with ISPA as fusion method. Thus, in these experiments, TriNet is used and trained as in Section 5.2.2.

For the view classification method of our pipeline (cf. Sec. 4.4), optimization is carried out using ADAM [Kingma & Ba, 2015], categorical cross-entropy loss, and learning rate of 0.0001 with a decay rate of 0.9 and 0.999. We found a convergence for MMV after 43 epochs. For FEP training purely, we found convergence after 37 epochs, and for MMV pre-training along with FEP fine-tuning, the model converged after 32 epochs.

5.5.3 Projection alignment

In this section we study qualitative results of projection alignment. Figure 5.12 shows one person on the left side in a multi-exposure FE image, whereas on the right side various aligned images are shown, which are automatically de-warped and aligned by the proposed algorithm. To process these images, the FE image sequence, together with bounding boxes, was fed into the projection alignment algorithm.

For most person locations in the FE image, the alignment worked well. One issue noted was that the accuracy of the bounding boxes not surprisingly influenced the result in a way which depends on the position in the FE image: a few pixels of misalignment next to the FE image center result in no visible error in the virtual image, whereas for the same error in the FE image with increasing off-axis displacement many more pixels in the virtual image contain background clutter (see in Figure 5.12 the clutter on the left and right of the person). Semantic segmentation in the FE image [Blott et al., 2018a] should be employed in future work to improve the bounding box accuracy by shifting person borders closer to image borders. Note, in this work we do not apply this step since we do not have such an abundance of training data for training the semantic segmentation network, see also the discussion in Section 4.2.

In general, our qualitative results from the 300-person FEP dataset confirm that the proposed algorithm can align a person in a virtual camera image if the person was initially oriented in an arbitrary pose and distorted by the FE projection.

5.5.4 Person view classification

In this section, the view classification approach with three interchangeable backbones (VGG16, ResNet50, and InceptionV3), mentioned in Section 4.4, is investigated using the two introduced datasets.

Four experiments are conducted for all classification approaches: i) Training and testing on the large MMV dataset to analyze the performance in case that sufficient training data is provided. ii) Training and testing on the new FEP dataset to evaluate the performance on a smaller dataset, which is more challenging due to the view-direction changing from nearly horizontal to more vertical. iii) Training on the large MMV dataset and testing on the FEP dataset to assess the generalization ability of the approach. iv) Training on the MMV dataset with fine-tuning on FEP training, and evaluation on FEP test dataset, as a realistic real-world scenario. We use the confusion matrix as the key performance indicator.

Tables 5.4 to 5.6 show the confusion matrices for the different network architectures. From left to right, the results are generated from: i) MMV training and testing (Tr-Te), ii) FEP Tr-Te, iii) MMV-FEP Tr-Te, and iv) training on MMV with fine-tuning on FEP training dataset and testing on FEP test dataset.

The findings are: performance on the large MMV dataset is in contrast to the FEP dataset nearly independent of the architecture; side views have the lowest classification rate, likely suffering from the fewer samples in the training set. The average overall classification accuracy is around 85%. Performance on FEP is generally better. The confusion matrices indicate that even with a relatively small training set, similar performance can be reached when pre-training on ImageNet is performed. Furthermore, the experiments indicate that there is still a large performance drop

Table 5.4: View classification with InceptionV3. Confusion matrices for training ("Tr") and testing ("Te") with different datasets. (GT = Ground Truth)

[%]	Tr: MMV; Te: MMV			Tr: FEP; Te: FEP			Tr: MMV; Te: FEP			Tr: MMV+FEP; Te: FEP		
GT	front	side	back	front	side	back	front	side	back	front	side	back
front	**85**	13	2	**78**	15	7	**76**	17	7	**91**	3	6
side	11	**77**	12	4	**84**	12	22	**70**	8	11	**74**	15
back	2	9	**89**	4	3	**93**	34	13	**53**	6	4	**90**

Table 5.5: View classification with ResNet50. Confusion matrices for training ("Tr") and testing ("Te") with different datasets. (GT = Ground Truth)

[%]	Tr: MMV; Te: MMV			Tr: FEP; Te: FEP			Tr: MMV; Te: FEP			Tr: MMV+FEP; Te: FEP		
GT	front	side	back	front	side	back	front	side	back	front	side	back
front	**89**	9	2	**84**	13	3	**59**	16	25	**93**	4	2
side	12	**76**	12	2	**89**	9	9	**50**	41	9	**83**	8
back	1	10	**89**	2	7	**91**	9	5	**86**	2	4	**94**

Table 5.6: View classification with VGG16. Confusion matrices for training ("Tr") and testing ("Te") with different datasets. (GT = Ground Truth)

[%]	Tr: MMV; Te: MMV			Tr: FEP; Te: FEP			Tr: MMV; Te: FEP			Tr: MMV+FEP; Te: FEP		
GT	front	side	back	front	side	back	front	side	back	front	side	back
front	**88**	10	2	**92**	7	1	**70**	12	18	**87**	11	2
side	11	**77**	12	12	**79**	9	21	**53**	26	11	**78**	11
back	3	9	**88**	6	2	**92**	23	6	**71**	5	4	**91**

Table 5.7: Accuracy (A), Precision (P), Recall (R), F1 Score (F1) , Overall Accuracy (OA), Average Accuracy (AA, average above the three classes) for the ResNet50. Values derived from Table 5.5.

	Tr: MMV Te: MMV				Tr: FEP Te: FEP				Tr: MMV Te: FEP				Tr: MMV+FEP; Te:FEP			
Class	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
front	0.92	0.86	0.88	0.87	0.93	0.95	0.84	0.89	0.81	0.77	0.59	0.67	0.94	0.89	0.93	0.91
side	0.86	0.80	0.77	0.79	0.90	0.82	0.89	0.85	0.76	0.70	0.50	0.58	0.92	0.91	0.83	0.87
back	0.91	0.86	0.88	0.87	0.93	0.88	0.91	0.90	0.73	0.57	0.86	0.68	0.94	0.90	0.94	0.92
OA/AA	85% / 90%				88% / 92%				65% / 77%				90% / 93%			

of around 18% by applying the MMV model on FEP without adaptation to target data statistic. However, by fine-tuning with FEP samples, the best performance on FEP again reaches around 90% with the ResNet50 architecture. In summary, we obtain a very promising classification accuracy on both datasets. The resulting view information can be used as strong prior knowledge for the next PRID module in our pipeline.

In general, the results indicate that the use of the ResNet50 gives the best results compared to the InceptionV3 or VGG16. Table 5.7 shows the extended measures for the ResNet50. The overall best accuracy for FEP is 90%, whereas for MMV 85% is obtained. Training with MMV and testing with FEP show a significant performance difference of -25% in overall accuracy compared to training with MMV along with FEP and testing with FEP.

As we assumed, the high performance of our data-driven classification method is due to training with samples from the target domain. Even though our projection-aligned-images look as taken by a central projection geometry cameras, the image statistics slightly differ, so that a model which is purely trained on central projection geometry images does not generalize well, and thus provides a lower performance compared to training with target domain images.

Besides the experiments indicate that for FEP a pre-training with MMV is useful (+2% accuracy). Furthermore, for MMV, the F1 scores indicate that front and back views can be distinguished equally well, but side views are worse. For FEP, it is also the same. Moreover, side views have the lowest recall for FEP and MMV, which explains why the F1 score, being the harmonic mean of precision and recall, is also low. However, for the experiment with domain shift (Tr: MMV, Te: FEP) these findings cannot be confirmed, since accuracy and precision are better for side views than for back views. However, the findings indicate that the data statistics is different between the two datasets.

For us, it is surprising to see that the overall accuracy of FEP is better than that of MMV, because the camera height of FEP is much larger and, thus, ambiguities in person shape increase. We believe a larger test dataset and a controlled environment is needed to investigate the reasons in the future.

Table 5.8: Multi-view PRID results. Multi-view PRID (MvP) @ X% view classification accuracy (VCA). Asterisk (*) indicates, the original dataset does not provide enough views.

Experiment / Dataset rank#1 [%]	**MMV**	**FEP**
Single-shot, random view per individual	84.8	50.7
Random-view fusion, no view awareness	94.0	60.3
MvP @ 30% VCA	93.8	59.9
MvP @ 40% VCA	93.8	63.7
MvP @ 50% VCA	94.3	64.1
MvP @ 60% VCA	95.2	66.9
MvP @ 70% VCA	96.3	68.5
MvP @ 80% VCA	97.0	**72.8**
MvP @ 90% VCA	97.4	*
MvP @ 100% VCA	**98.1**	*
MvP @ ResNet50 (highest confidence)	97.0	72.0
MvP @ ResNet50 (random confidence)	96.2	70.0

5.5.5 Assessment of PRID results

In this section, the influence of the individual components on the final PRID performance of our pipeline is analyzed first; the results of which are presented in Table 5.8. As a reference, the re-mapped Market dataset (MMV) which was not recorded by the FE lens is used. For all FEP experiments, a pre-training based on the ImageNet dataset [Deng et al., 2009], MMV dataset and fine-tuning on FEP training dataset were performed to pre-train the weights on large datasets.

Rank#1 accuracy is reported, this is the probability of finding the correct match in the first rank of the Cumulative Matching Characteristic (CMC) curve [Gray et al., 2007] (cf. Sec. 2.3).

Effect of multi-view PRID and view classification accuracy (cf. 3-6 of Fig. 4.4): If the single-shot strategy is applied, as commonly used approaches do, a rank#1 performance of 84.8% for MMV and 50.7% for FEP is obtained. Note that the datasets (MMV and FEP) consists of many more views per person than two classical cameras typically provide. Thus, more views are available and it is easier to find a corresponding match. Consequently, a higher re-identification performance is expected. By carrying out fusion with random-view elements from the *3-image views*, i.e. images without a view attributes, rank#1 accuracy is improved to 94% for MMV and 60.3% for FEP. For multi-view PRID, different sets corresponding to different view classification accuracies (VCA) ranging from 30% to 100% are sampled from the ground truth view annotations. This means, for each 3-view set, the probability that an element consists of the correct view is simulated to study the influence of view classification quality. In other words, for each 3-view set ($\mathbb{S}_\mu = [\mathbf{F}, \mathbf{B}, \mathbf{S}]$, cf. Sec. 4.1) the probability that e.g. $\mathbf{F}$ really shows a front view is VCA%. The probability that $\mathbf{F}$ shows a back or side view is consequently

Figure 5.13: Illustration of different image quality; Top: frequently cited Market-1501 dataset (Superset of MMV). Bottom: FEP after projection alignment.

100%-VCA%. The same is applied to **B** and **S** independently of each other. Thus, we can analyse the robustness of multi-view PRID against view classification failures. Note, since the sampling of the 3-view sets is a random process in the selection of IDs with a wrong view classification, we employ cross-validation, evaluate ten dataset splits per VCA and average the rank#1 results.

Obviously, the performance of multi-view PRID increases with increasing VCA. The best rank#1 is 98.1% on MMV with 100% VCA and 72.8% on FEP with 80% VCA. Higher accuracy for view-classification did not provide enough views in the underlying FEP dataset to obtain meaningful results. This is because for a VCA of, e.g. 90% we need, across all sets in each element, 90% of correct views. However, not all persons are available in all views (cf. Sec. 5.5.1). Consequently, when only 80% of persons are available in front, side, and back views, the upper VCA border is 80%.

Interestingly, for MMV the last 20% view classification errors only have a minor influence on the overall results, decreasing the performance by only 1.1%. This means that the multi-view PRID tolerates a certain amount of view classification errors, which is a very interesting property for applications in the real world.

On the Market dataset, which has a very high baseline performance (94.0% for random-view fusion), the improvement of the multi-view approach amounting to 4.1% is still significant; for FEP the improvement is 12.5% and thus considerably larger. The comparably lower rank#1 performance for FEP is probably caused by the smaller training dataset, and much steeper person views as illustrated in Figure 5.13. The experiments indicate that the multi-view approach for PRID outperforms the re-identification with single-shot and random-view fusion by a large margin of +13.3% /+4.1% rank#1 for MMV and +22.1%/ +12.5% rank#1 for FEP.

Combined evaluation of view classification and matching: In the last two lines of Table 5.8, the results are depicted for PRID using the ResNet50, which showed the best performance in

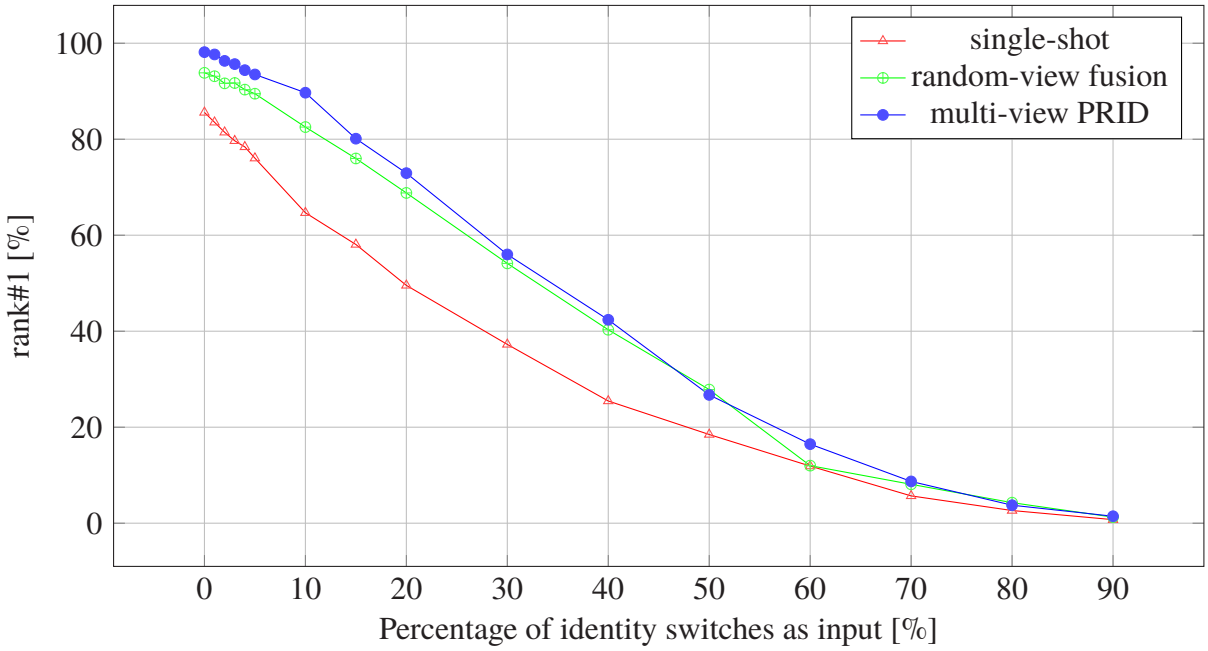


Figure 5.14: Influence of identity switches. Identity switches from 0% to 90% are simulated per 3-view set element and rank#1 is reported for the MMV dataset.

Section 5.5.4 for view classification instead of a simulated classification result. The obtained performance for the classification result with the highest score is 97.0% for MMV and 72.0% for FEP, thus, close to the simulated view classification accuracy which underlines the effectiveness of the proposed re-identification approach. In the last line of the table, the results for PRID with a random confidence score for particular views in view classification are presented. This means instead of using the sample of highest confidence for a particular view in a 3-view set, a random sample is selected which is also classified as the particular view. The results indicate that using the highest confidence gives better performance, even though these samples could be overconfident.

Effect of identity switches (cf. 1-6 of Fig. 4.4): In real-world applications, tracking results instead of ground truth for person detection, and thus data of inferior quality, must, of course, be used as input for person re-identification. By way of example, we present the results of the effect of ID switches, a common problem during tracking. Such switches result in 3-view sets being composed of different persons instead of only showing one and the same. Different levels of ID switches are simulated and evaluated for the MMV dataset to study their influence on the person re-identification performance. In other words, for each 3-view set ($\mathbb{S}_\mu = [\mathbf{F}, \mathbf{B}, \mathbf{S}]$, cf. Sec. 4.1) the probability that, e.g., **F** shows really ID μ is simulated. The probability that **F** shows a different person is ν%. The same is applied to **B** and **S** independently of each other. Thus, we can analyse the robustness of multi-view PRID against identity switches.

Figure 5.14 depicts the results. Not surprisingly, single-shot is not robust against these ID switches, whereas multi-view PRID and random-view fusion tolerate certain errors. As can be seen, multi-view PRID performs better than random-view fusion. However, the performance

difference decreases with an increasing number of ID switches. As a reference, the current best tracker, *ISE_MOT*, in the MOT-2017-Challenge score leader-board[16] has 2,389 identity switches in 2,355 trajectories in total. The tracker with the fewest switches, *TT17*, shows 1,088 identity switches but a lower key performance indicator (*MOTA*) for person tracking. It is thus clear that tracking errors are another performance barrier for a complete person re-identification pipeline. With the current state-of-the-art tracking approach on perspective cameras, the advantage of the proposed multi-view PRID is less visible compared to random-view fusion than with perfect tracking results. Consequently, better tracking is a mandatory prerequisite to exploit the full potential of multi-view PRID. As we take images from the nadir direction and thus have much fewer person occlusions, we expect to be able to reduce the number of ID switches significantly compared to horizontal views. Due to fewer occlusions, the association between persons in crowded scenarios should be improved, since technique such as optical flow could be used to track a person in a scenario.

5.5.6 Comparison with a contemporary approach

Following the extensive evaluation of design choices, expected weaknesses, strengths, and the respective pipeline components, in this section our approach is evaluated and compared to the closest related work [Eberle, 2018].

Overview: The goal of this experiment is to demonstrate that the PRID performance for our modular pipeline hardly improves further when more training data is used; whereas an end-to-end learning approach shows a low PRID performance by using the same training data. Thus, the end-to-end learning approach would benefit from much more data to show better performance. However, more data for end-to-end learning is currently not available to us and, from a practical point of view, expensive to provide. This is because, i) more target domain images have to be recorded using multiple cameras, ii) annotations of persons including a global assignment are needed, and iii) the General Data Protection Regulation (GDPR) makes it challenging to record persons in public spaces within the European Union.

As there is no previous work using multi-view observations for PRID along with fisheye cameras, we extend [Eberle, 2018], which is the state-of-the-art for view specific PRID (cf. Sec. 2.6), for our multi-view purpose, and design particular experiments for fair performance comparison.

Specifically, steps (3) to (5) of our pipeline (cf. Fig. 4.4) are replaced by an approach where view classification, re-identification, and fusion are applied in one common network architecture. Consequently, there is no "classical" input-output end-to-end learning, since the fisheye effect and arbitrary person orientations are eliminated by our proposed projection alignment approach, and not by some additional layers which might learn a similar mapping. Rather, view classification,

[16] `https://motchallenge.net/results/3D_MOT_2015/?chl=3&orderBy=MOTA&orderStyle=DESC&det=Public`, accessed on December/31/2019

feature extraction, and fusion are done in an end-to-end fashion, while input is provided by our proposed alignment approach.

The baseline: To fairly conduct the experiments, we use *Pose Sensitive Embedding* (PSE) [Eberle, 2018] in the extended version called *Multi-view Sensitive Embedding* (MSE) [Salam, 2019]. Eberle proposes an architecture which is able to classify the view, the person pose, or both, and use them as prior knowledge for PRID. In the following section we focus on the view part only (one module of Figure 5.15) where a network-branch is trained to classify the view (front, back, or side), and particular feature maps, so-called *Base Re-Id Feature Maps*, are extracted to obtain a *view specific embedding* for PRID. The final embedding is extracted with respect to the classified view, since depending on the view-classification-score, a weighting scheme is applied to activate the corresponding so-called *View Unit*. Such a unit is basically multiple network layers that are learned for a particular view. Thus, the input to the network is a single image, and the output is the corresponding embedding. That means, feeding, e.g., a front view into the network, should be ideally classified as a front view, and the view unit for a front view is activated to extract a front-view-embedding.

In contrast to our modular pipeline, i) the approach is trained in an end-to-end fashion ((3) to (5) of Fig. 4.4 are one network), and ii) matching of arbitrary view combinations is pursued. Thus, MSE has another goal as our pipeline has. Because we want to systematically match the same views to improve issues stemming from intra-person and inter-person variation. Thus, Eberle performs a view-aware matching without constraining view-combinations, which enables a general matching. However, it can suffer from the large intra-person variation caused by arbitrary view matching. Thus, for a fair performance comparison, we design the following adapted baseline and experiments.

The adapted baseline: The main idea of MSE [Salam, 2019] (cf. Fig. 5.15) is to feed three images per person and camera into a network with one front, back and side view in a random order, let the network itself decide which view is seen[17], and how to weight feature maps to calculate a common embedding. In other words, a 3-input-images–to–1-output-embedding approach is proposed.

To realize this approach using the approach of Eberle, PSE with the configuration for view is employed as a module, i.e. no pose information is exploited. Then, three modules are used, one for each input image, whereas the learned network weights are shared across the modules. Afterwards, a 1536-dimensional fully-connected layer[18] as in [Eberle, 2018] is employed as a fusion layer to the feature maps to obtain the embedding with respect to the classified view and the *Base Re-id Feature Maps*. The training in end-to-end fashion can be summarized as follows: i) The ResNet50 backbone weights are initialized with weights that are obtained from a pre-training on the ImageNet dataset [Deng et al., 2009]. ii) The view classification is learned

[17] A *view predictor* in [Eberle, 2018] and [Salam, 2019], is called view classification in our approach.
[18] Note, there is no argument for using *1536*, but [Salam, 2019] showed that this gives the best performance

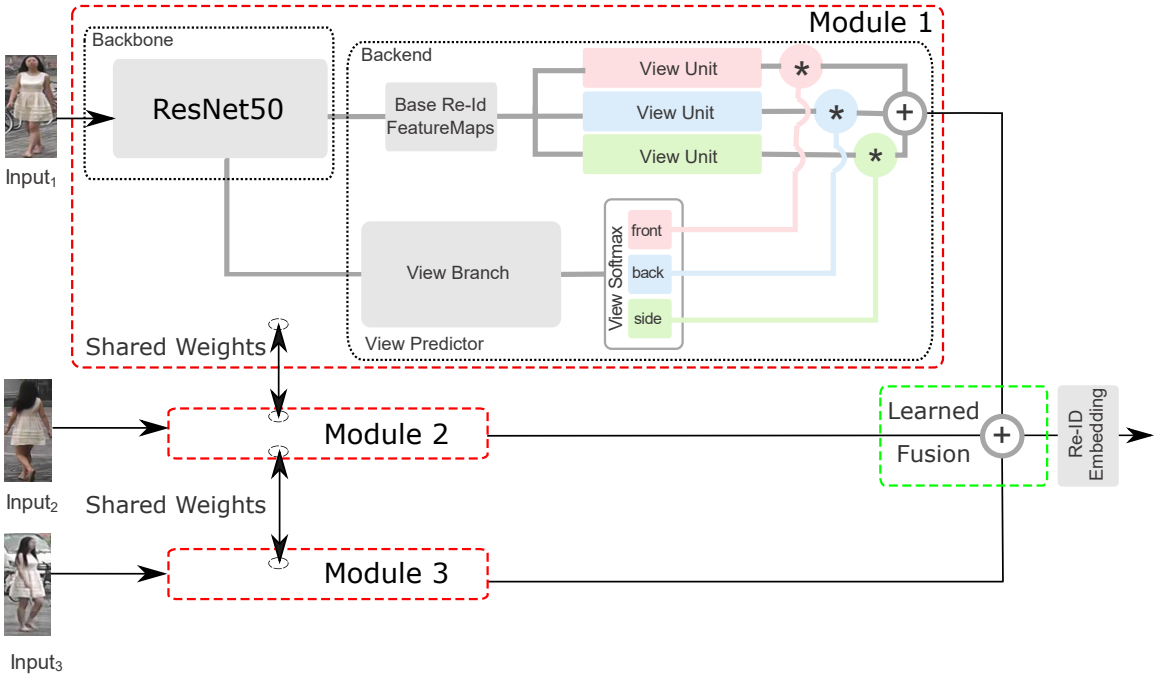
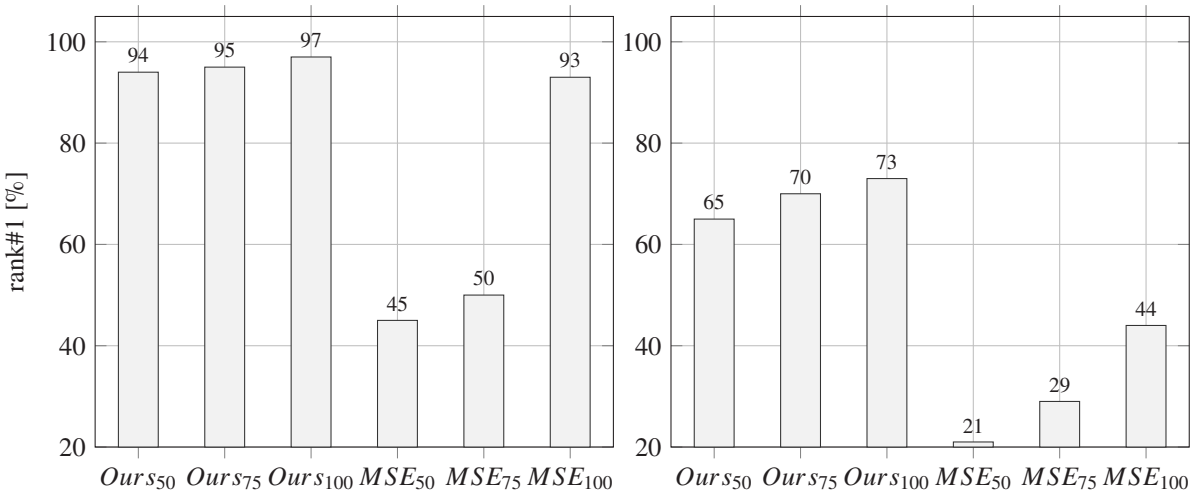


Figure 5.15: MSE overview, adapted from [Salam, 2019]. A module is equivalent to the contribution of [Eberle, 2018]. MSE extracts, for three input images of one person, a feature vector as output. The view prediction, re-Id features per view, and fusion are in learned end-to-end fashion.

by freezing the re-id layers. iii) The re-identification is trained after the view classification was trained. iv) After the 1-input re-identification was trained, the three modules and the fusion layer are trained for our goal, 3-input-images to one output embedding. v) Hyper-parameters of the architecture, loss function (*classification loss*), training strategy, and implementation details for MSE are set according to [Salam, 2019], as they also show the best performance on the evaluated datasets.

Dataset and experiment design: The fisheye dataset (FEP) and the re-mapped Market dataset (MMV) are used. The input of three images consists of one image per view, and is sampled once from ground truth to obtain three different views for each person per camera. In case a specific view for a particular person is not available, the element is replaced by another randomly selected view. We use the input of three images since MSE does not contain a sampling strategy and cannot deal with sequences as our approach can. Consequently, we have to define and select the network input data in advance and feed exactly the same samples into both approaches for a fair performance comparison. For all experiments on the FEP dataset, the network is pre-trained on MMV before training on FEP is applied, precisely as in our experiments above, see Section 5.5.5. Note that for learning, both approaches use exactly the same training images and annotations.

For all experiments the respective test dataset is then fixed (MMV or FEP), and three networks are trained with [50%, 75% 100%] of available training IDs, for each approach, respectively. Consequently, MMV has 269 persons for testing, and [467, 700, 934] persons are used for the training. Besides, FEP consists of 100 persons for testing and [100, 150, 200] persons are



(a) MMV Dataset (b) FEP Dataset

Figure 5.16: Results of conducted experiments. (a) MMV Dataset, (b) FEP Dataset.

available for the training. This setup allows us to compare the influence of using more IDs for training for each dataset, whereas the respective test dataset is fixed.

Results and discussion: The results of the conducted experiments for both datasets are depicted in Figure 5.16. In general, the results indicate that our approach performs better than MSE (MMV: +4% rank#1 by using 100% training data; FEP: +29% rank#1 by using 100% training data), even though PSE performs better than the TriNet [Hermans et al., 2017] in the original publication on the Market-1501 dataset. Again, the TriNet is merely the feature extraction method employed in our modular pipeline. More importantly, there is a large performance difference between ours and MSE when fewer IDs are used for training. This confirms our findings of Section 5.4.3 where we found a performance saturation for the TriNet.

In the following paragraph we discuss the MMV dataset results more in detail: for our approach, rank#1 is increased by +3% by using twice as many IDs for training (50% to 100% IDs for training, i.e. +467 persons), whereas for MSE, we find +48%. This underlines the need to use much more training samples for MSE whereas our approach indicates a low performance gain by using more images. We find the same behavior for the FEP dataset: for our approach rank#1 is increased by +8% by using twice as many IDs for training (i.e. +100 persons), whereas for MSE +23% are found. We believe the large performance gap compared to MMV shows that FEP is much more challenging than MMV, and, thus, many more samples are needed to tune the networks to the different data statistics. Additionally, the experiments indicate that the modular pipeline better bridges the domain gap from MMV to FEP. Whereas the performance difference between our approach and MSE is 4% rank#1 on MMV, the performance difference is 29% rank#1 on FEP.

Experiment conclusion: The experiments and discussions confirm the results of the experiments conducted in Section 5.4, where a performance saturation was found for the TriNet, even though no modular processing was applied and only TriNet was employed. In conclusion, this

indicates that for a small dataset a modular processing can perform better than an end-to-end learning approach. However, we clearly state that the lower performance of MSE is likely due to the small training dataset. Thus, the approach could improve the PRID performance by using more training data and also perform better as a modular pipeline. However, creating a large scale dataset for end-to-end learning is expensive, whereas a modular pipeline seems to need fewer training samples to provide a high performance for the given datasets.

Obviously, the results are an interesting finding, because both approaches (our approach and MSE) employ a ResNet50 as a backbone with weights that are obtained by a pre-training on the ImageNet dataset. In addition, we use i) the categorical cross entropy loss to train the view classification with a first network, ii) the triplet loss with hard-negative mining for the person re-identification for a second network, and iii) a rule-based fusion. MSE uses i) the classification loss, ii) no hard-negative mining, and iii) a learning-based fusion. Thus, one network is trained for all. We therefore believe, the performance differences will mainly appear due to the applied loss-function in combination with the backend of the network. Probably, the hard-negative mining is essential to train with a small dataset like FEP.

Besides, there is another advantage in applying a modular pipeline. One assumption for within-video multi-view is that a person tracking is applied to obtain multiple images of the same persons. By using the tracking information to obtain the view (cf. Sec. 2.6) and fusing it with our deep learning based view classification to an ensemble of classifiers, both approaches could easily be integrated into our approach. A replacement of our view classification with an ensemble of independent approaches can help to improve the PRID performance further, whereas the integration of an ensemble into MSE seems to be more difficult due to the selected architecture. Hence, MSE learns the view classification and the feature extraction from statistics, whereas we could easily integrate an independent view classification, which can help to better generalize unseen open-world data-statistics.

5.5.7 Qualitative comparison

After the quantitative performance evaluation of person re-identification, qualitative impressions are presented to underline the quality of the results of our multi-view strategy. We use the FEP dataset of Section 5.5.1, and we show examples where a strategy without multi-view fails, whereas our multi-view strategy shows a much better person re-identification performance. Moreover, we give an example and discuss situations where our approach did not find the correct match. Figure 5.17 and Figure 5.18 show person re-identifications of six different probes (one per double-row).

The odd rows (single-shot) depict the corresponding top-10 matches for the FEP dataset by using one image per view per camera[19], matching between Camera_2 to Camera_0, and applying the

[19] i.e. 1×front, 1×back, 1×side image per individual is in the gallery

single-shot strategy. This setup is unusual compared to the previous experiments, but allows us to directly observe the differences between the top-10 images compared to the probe, while in parallel the ranked images can be seen.

The even rows (multi-view) show the results of the same test dataset when our multi-view strategy is applied, 3-view sets are used, and the same network and weights are employed to extract feature vectors. Further, only rank#1 and rank#2 are depicted.

The figures indicate that the single-shot strategy cannot find the depicted samples. We assume our dataset is too difficult compared to public datasets, which have low camera heights; persons with highly asymmetric appearance are also rare in the public datasets. However, by using our proposed strategy, much more correct matches are found.

By analyzing the issues of matching without view-awareness, one can observe a highly asymmetric person appearance which is one major problem. The most miss-matches are probably due to: i) open jackets (examples: ①, ④), which change appearance if a view of different direction is provided in the gallery. ii) front logos on shirts (example: ②), which increase ambiguities depending on the logo size. iii) carried objects (examples: ③,④), which are not independently modeled in the re-identification approach; rather, the appearance is integrated regardless of an auxiliary-class. The last row shows an example where our multi-view strategy failed (example ⑥). The 3-view set of the probe shows the dark-blue color of the jacket and a light-blue color for the shirt of the person, whereas for the correct gallery sample the same jacket looks black, and the shirt looks white. Thus, we believe the reason for the wrong match is that the network did not learn the color transfer from the camera used as a probe to the camera used as a gallery. We analyzed the dataset in detail and found multiple samples that show a color transfer from one camera to the other. Thus, we believe a proper photometric camera calibration to be beneficial for person re-identification to achieve color fidelity; otherwise, a large number of training samples are needed to teach for each target camera "*handovers*" to bridge the gap between the same colours in different cameras. However, proper photometric camera calibration could be challenging in the security camera domain. This is due to automatic camera control, which automatically adjusts the gain, exposure, aperture, and white balancing to fit for a bright day scenario and a low light night scenario.

Our proposed multi-view approach, therefore, in general, can be seen as a strategy that helps to boost the re-identification performance further, whereas the advantage of an image-based modality is preserved, and the employed feature extraction method can be seen as an interchangeable module.

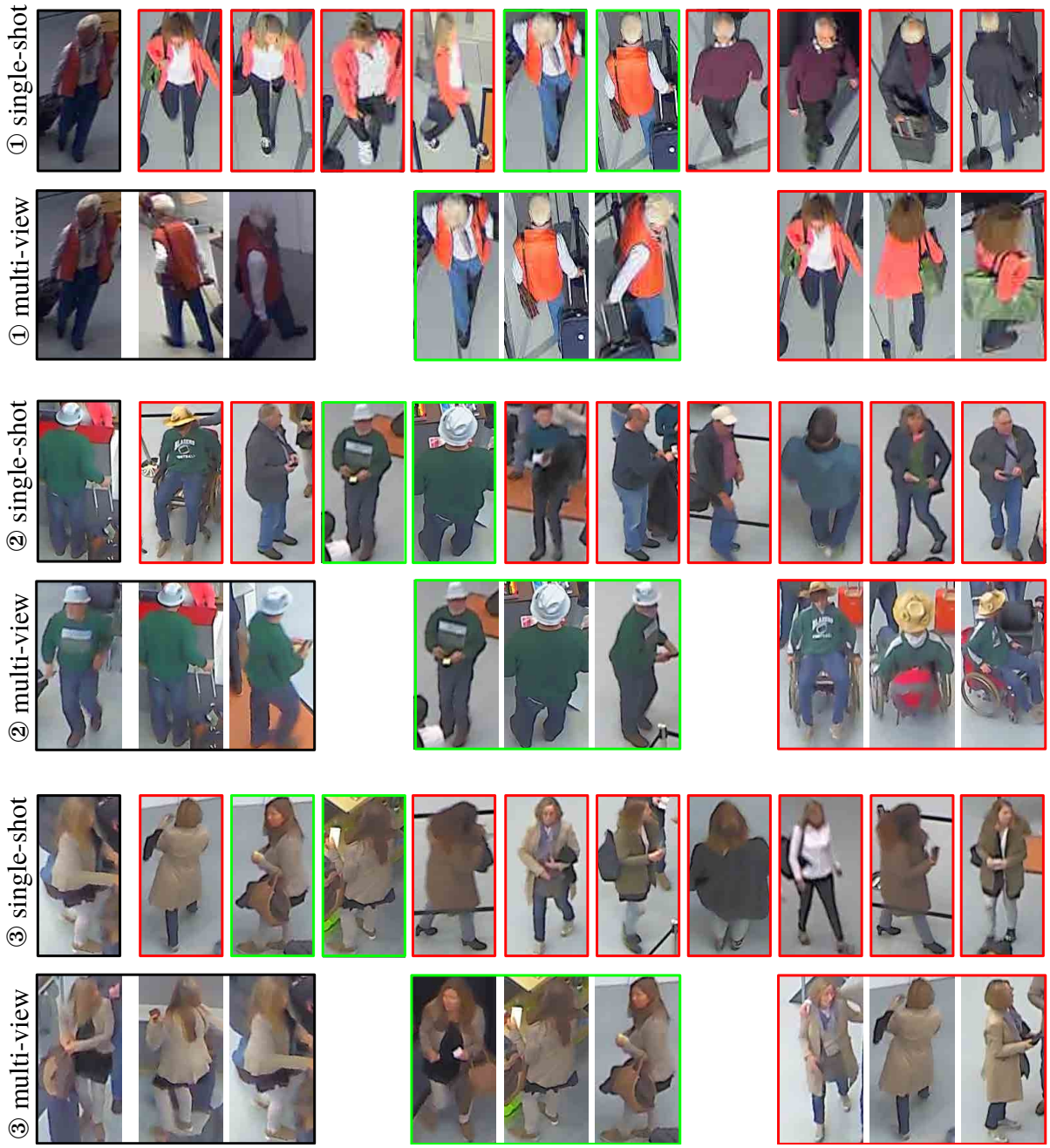


Figure 5.17: Qualitative PRID results (1/2). Three different person matchings for the FEP dataset are shown. Odd rows show matching with single-shot strategy. On the left (black border) the probe is depicted. To the right the ten matches of decreasing similarity are shown. Correct matches in green bounding boxes, false matches in red bounding boxes. Even rows show matching with multi-view strategy. On the left the 3-view set of the probe is depicted. On the right the rank#1 and rank#2 3-view sets are shown. It can be seen, the correct match is found by our view-aware approach even though highly asymmetric person appearance is visible.

Figure 5.18: Qualitative PRID results (2/2). Three different person matchings for the FEP dataset are shown. Odd rows show matching with single-shot strategy. On the left (black border) the probe is depicted. To the right the ten matches of decreasing similarity are shown. Correct matches in green bounding boxes, false matches in red bounding boxes. Even rows show matching with multi-view strategy. On the left the 3-view set of the probe is depicted. On the right the rank#1 and rank#2 3-view sets are shown. It can be seen, the correct match is found in 2/3 of the illustrated cases by our view-aware approach even though asymmetric person appearance is visible.

Chapter 6

Conclusions and future work

This chapter summarizes the results of this thesis and gives an outlook on future research directions.

Our research goal was the improvement of appearance-based person re-identification for the security camera domain, by introducing multi-view information and integrating it as strong prior knowledge in the re-identification process. Additionally, we wanted to propose an approach to obtain purely image-based multi-view information with a single camera. The goals were both achieved by a novel multi-view approach.

To evaluate our proposed multi-view approach, comprehensive experimental investigations were conducted which revealed the following:

- Using fisheye cameras in nadir pose provides a much wider variety of person-observations in the respective camera. More person views are available compared to a classical camera. Besides, this work also provides useful insights for scenarios where persons are guided, e.g., with separation tape, in combination with central projection cameras, to obtain multi-view observations. (cf. Sec. 5.5)
- Multi-view person re-identification outperforms standard multi-shot matching by a large margin (up to +18% rank#1 in our experiments), benefiting from better handling of intra-person variation and inter-person variations. (cf. Sec. 5.2)
- Multi-view person re-identification as a strategy is independent of feature design, does not need a very large amount of training data, and can boost the re-identification further, even though a certain performance saturation is observed in the feature extraction method. (cf. Sec. 5.4)
- In a further analysis, it is shown that detailed image information hardly improves the PRID performance. Moreover, our experiments indicate that person head information is barely considered for re-identification, and there is rarely a performance drop if same person

images are blurred. Although image quality deteriorates by our hardware setup, from persons' shoes to the head, this blurring does not negatively affect the re-identification. (cf. Sec. 5.4)

- The novel hardware setup allows the provision of high-resolution bird's eye views of persons whenever persons pass the central image region. However, experiments show a poor performance due to the increasing ambiguities, which are typical for a bird's eye view. Additionally, due to the narrow central image region of a fisheye camera, bird's eye views are hard to obtain. Therefore, conducted experiments indicate to omit bird's eye views of persons until pose-invariant feature extraction methods are available. (cf. Sec. 5.3)

In general, our modularized software pipeline allows the use of almost any state-of-the-art feature extraction method regardless of the architecture and training loss function. Indeed, re-ranking post-processing methods, which are not addressed in this work, could further improve the performance, as many more images are available of persons by recording with higher frame rates. However, this comes at the expense of a shorter exposure time, higher data-bandwidth, and more computational costs. Therefore, the lower performance bound of our proposed approach is given by the feature extraction method, including the quality of person detection and tracking. Moreover, the re-identification performance difference between a classical camera and the multi-view fisheye approach supposingly becomes larger with increasing dataset size and level of asymmetric person appearance within the dataset, because a classical central projection camera approach is typically not able to handle highly asymmetric person appearance.

Finally, one question arises: is our novel approach ready for person re-identification systems to be used in practical applications? This can be answered from different aspects. Person re-identification is not ready for fully autonomous systems, but still applicable for use in some cases of real-world scenarios, e.g. shops which can be described as a closed-set scenario. The re-identification is modeled as a retrieval task, and thus a human operator could be relieved at work by our approach since the probability of finding the correct match within the top ranks is increased. Currently, human operators using person re-identification systems have to consider a lot of dissimilar persons. Thus, we could reduce the possible search-space by robustly eliminating dissimilar candidates to save working time and human effort.

There are several ways to extend the work of this thesis regarding future research.

The task of person detection and tracking in the fisheye images is not the focus of this work, but assumed to be perfectly available. Concerning practical applications, it is essential to develop a sound solution and to evaluate the PRID performance in combination with person detection and tracking at system level.

Our view classification is categorized into front, back, side, and bird's eye views of persons. Given an end-to-end learning architecture and a hierarchical clustering algorithm, deciding which classes to use could be helpful. This is due to the assumption that the representation learned from data statistics could differ from the one pre-defined by humans. However, a large dataset, which is currently not publicly available, is a mandatory prerequisite for conducting such experiments.

As discussed in Section 4.4, the view classification is a data-driven method. Using intra-camera tracking and moving directions of persons for a second view classification method, which is independent of the data-driven view classification, and fuse both methods to an ensemble, will likely boost view classification accuracy, and thus the person-re-identification performance, further.

Another interesting use-case for large scale dynamic person re-identification is the suspension of the fisheye camera with nadir orientation under unmanned area vehicles (UAVs) of a *"flying camera network"* to allow for person re-identification outdoors. Here dynamic background and different illuminations occur which are challenging for PRID.

As an extension of this thesis, cameras equipped with panomorph lenses [Thibault, 2010] should be studied. These are panoramic lenses which have controlled distortions and thus show region depending magnification in ultra-wide-angle image space, which could lead to a higher resolution of persons by using the same sensor. Furthermore, they allow the analysis of more details compared to a fisheye lens in relevant regions. In contrast to our findings in Section 5.4.3, where we showed that detailed image information is not considered, another feature extraction method is mandatory which can handle these details.

Finally, a data-efficient domain adaption technique is needed to address the generalization ability. This ability is vital to apply a trained network also in other scenarios where the camera, camera pose, background, or style of clothes differ. Furthermore, it is also important to evaluate the person re-identification performance with smaller network architectures to determine how many parameters are needed to successfully re-identify persons, e.g., on embedded-edge-devices. Another interesting aspect, from a scientific point of view, is to investigate, what is the smallest appearance-difference between two imaged persons that allows finding the correct match. Can this difference be expressed somehow by the hyper-parameters of a CNN architecture such as convolution size, stride, and pooling strategy?

Appendix A

Datasets

A.1 Our novel datasets

For the underlying research, **11** datasets are created, which consist of nine new sets created from scratch and three dataset re-mappings. Depending on the particular research goal individual sets consist of:

- pixel-wise class annotation for semantic segmentation [Blott et al., 2018a].
- an airport scenario dataset, consisting of three central projection cameras, 85 persons, a human bounding box annotation and global person ID association [Blott et al., 2018b].
- a second airport and shop scenario consisting of five fisheye cameras, 300 persons, human bounding box annotation and global person ID association [Blott et al., 2019].
- three different datasets for bird's eye view scenarios. We provide: fisheye projection and central projection.

Table A.1: Datasets underlying this thesis and research

Experiment	**New datasets**	**Re-mapped public datasets**	Σ
Semantic Segmentation [Blott et al., 2018a]	3	1	4
View-Aware PRID [Blott et al., 2018b]	1	1	2
FE PRID [Blott et al., 2019]	1	1	2
Bird's Eye Views Experiments	3		3
Σ	8	3	11

A.2 Public datasets

With the rise of PRID publications, different datasets were proposed for performance evaluation (cf. Fig A.1). Early works such as VIPeR [Gray et al., 2007] contain a small amount of data (632 individuals and 1264 images), annotated and cropped by humans.

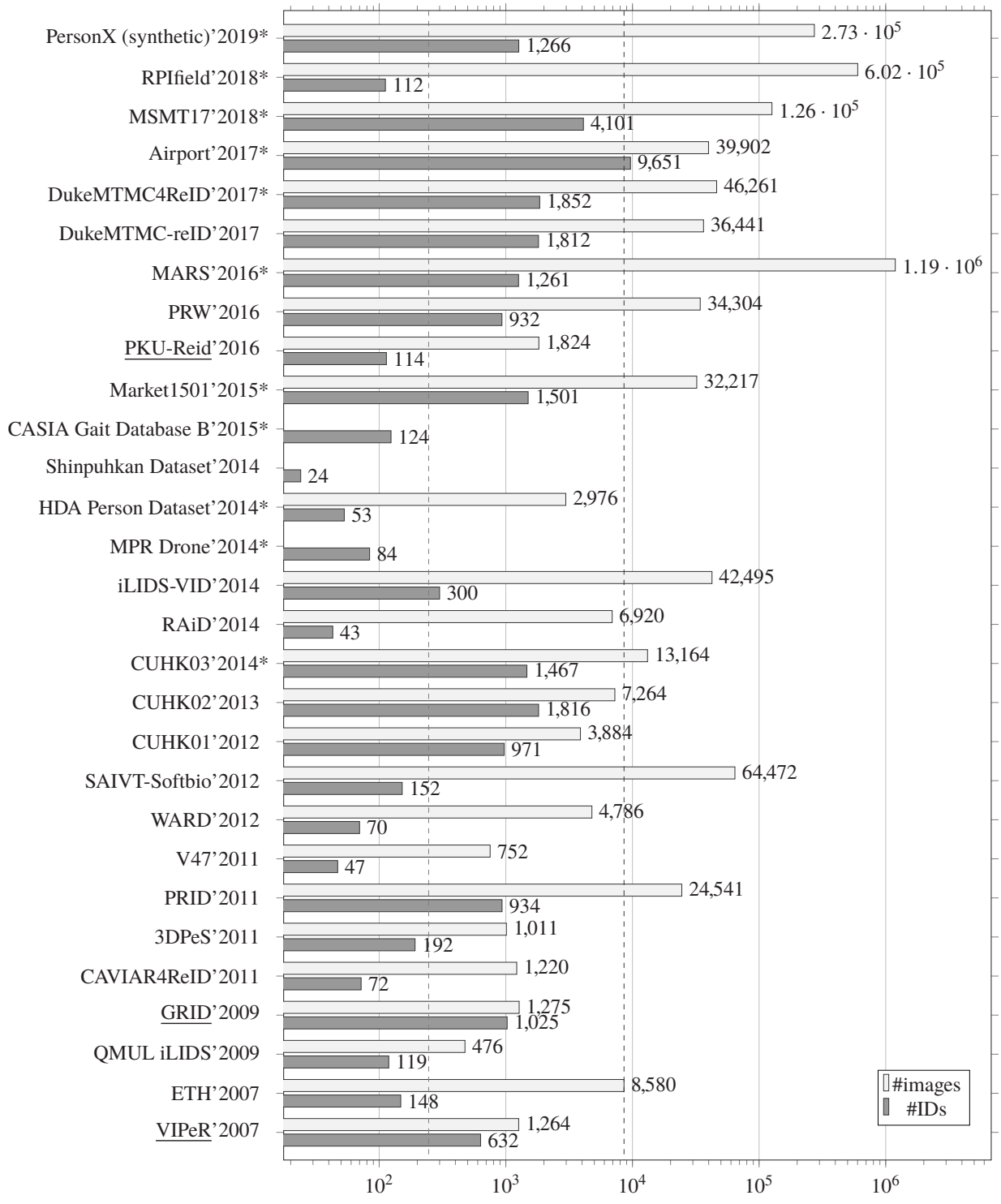


Figure A.1: PRID dataset overview: the dataset name, publication year, number of IDs, and number of images are shown. Underlined datasets merely provide single-shot images; the others provide multi-shot. * indicates person detection was automatically applied, others have manual annotation. For MPR Drone, Shinpuhkan Dataset, and CASIA Gait Database B the number of images is not known. Furthermore, the two dashed lines depict the median number of IDs and images, respectively. Note, the axis of abscissae has a logarithm scale. For up-to-date-references to the datasets, see footnote 1.

Larger datasets, such as Market-1501 [Zheng et al., 2015a], were mandatory with deep learning based strategies, which learn based on a huge amount of data statistics (1501 individuals, 32217 images). Here the gallery persons are detected by a deformable part base model approach, an outdoor campus scenario is recorded, and the probe persons are annotated by hand. However, the heights of the six cameras seem low (we assume a tripod was used and camera height is approximately at eye level), and persons are unsurprisingly horizontally projected into the images. The camera pose is close to an egocentric view [Fergnani et al., 2016] where the whole person is captured. This gives a strong prior knowledge for the re-identification and does not allow for performance comparison with arbitrary camera poses, e.g., security camera poses. While *Market-1501* is designed for single-shot, the *Mars* dataset [Zheng et al., 2016a], which is created from the same recording session as Market-1501, is created for video-based person re-identification evaluation. Here, continuous time frames (videos) are exploited for PRID but there are less IDs available than in the Market-1501 dataset. The *DukeMTMC* dataset [Ristani et al., 2016] and derivatives were published (1812/1582 individuals, 36441/46261 images). Again, a tripod was probably used for data capture, and, therefore, experiments with this dataset are not directly comparable to different camera poses. Other much larger datasets were published by the community.[1] The usage of egocentric view data mainly has the advantage that, e.g., the person pose, a good semantic segmentation, and the prior knowledge in which parts of the cropped bounding box particular body part lie can be learned by using publicly available datasets from other domains. However, concerning security camera installation points, where cameras are mounted in a height of three meters and more along with a large pan angle, to the best of our knowledge, hardly any public datasets exist and state-of-the-art strategies, e.g. obtaining the person skeleton, inherently fail, due to missing training data.

CUHK03 is another re-identification dataset that is frequently used to evaluate PRID performance. CUHK03 consists of 1467 IDs from ten cameras (5 pairs) and 13164 images. Again, a tripod was probably used for data capture.

The most employed and processed datasets in the era of deep learning are the datasets: Market-1501 [Zheng et al., 2015a], Duke [Ristani et al., 2016] and CUHK03 [Li et al., 2014]. In Market, the persons mostly wear short trousers, and the scenario looks like it was recorded in summer. In the Duke dataset the persons wear coats and long trousers so it was likely recorded in spring or autumn. In CUHK03, the persons were probably imaged between spring and autumn, as no winter clothes are visible. An highly asymmetric person appearance is underrepresented within the datasets and an open-world dataset consisting of various seasons, scenarios, ethnic groups, and camera poses is lacking. Most datasets typically look like the one depicted in Figure 2.3.

[1] `http://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html`

References

[Abraham & Förstner, 2005] Abraham S, Förstner W (2005) Fish-Eye-Stereo Calibration and Epipolar Rectification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59: 278–288.

[Aghajan & Cavallaro, 2009] Aghajan H, Cavallaro A (2009) *Multi-Camera Networks: Principles and Applications*. Orlando, FL, USA: Academic Press, Inc.

[Alkoot & Kittler, 1999] Alkoot FM, Kittler J (1999) Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20 (11-13): 1361–1369.

[Amari & Nagaoka, 2000] Amari Si, Nagaoka H (2000) *Methods of Information Geometry*, volume 191.

[Andriluka et al., 2014] Andriluka M, Pishchulin L, Gehler PV, Schiele B (2014) 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 3686–3693.

[Arsigny et al., 2006] Arsigny V, Fillard P, Pennec X, Ayache N (2006) Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices. *SIAM J. Matrix Analysis Applications*, 29 (1): 328–347.

[Bak et al., 2018] Bak S, Carr P, Lalonde J (2018) Domain Adaptation Through Synthesis for Unsupervised Person Re-identification. In: *European Conference on Computer Vision (ECCV)*: 193–209.

[Bak et al., 2014] Bak S, Zaidenberg S, Boulay B, Brémond F (2014) Improving Person Re-identification by Viewpoint Cues. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*: 175–180.

[Barreto & Araújo, 2001] Barreto JP, Araújo H (2001) Issues on the Geometry of Central Catadioptric Image Formation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 422–427.

[Bazzani et al., 2010] Bazzani L, Cristani M, Perina A, Farenzena M, Murino V (2010) Multiple-Shot Person Re-identification by HPE Signature. In: *International Conference on Pattern Recognition (ICPR)*: 1413–1416.

[Bishop, 2007] Bishop CM (2007) *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer.

[Blott & Heipke, 2017] Blott G, Heipke C (2017) Bifocal Stereo for Multipath Person Re-Identification. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W8: 37–44.

[Blott et al., 2018a] Blott G, Takami M, Heipke C (2018a) Semantic Segmentation of Fisheye Images. In: *European Conference on Computer Vision, Workshops (ECCV-W)*: 181–196.

[Blott et al., 2018b] Blott G, Yu J, Heipke C (2018b) View-Aware Person Re-identification. In: *German Conference on Pattern Recognition (GCPR)*: 46–59.

[Blott et al., 2019] Blott G, Yu J, Heipke C (2019) Multi-view Person Re-identification in a Fisheye Camera Network with Different Viewing Directions. In: *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*

[Bromley et al., 1993] Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R (1993) Signature Verification Using a Siamese Time Delay Neural Network. In: *Neural Information Processing Systems (NIPS)*: 737–744.

[Brown, 1966] Brown DC (1966) Decentering Distortion of Lenses. *Photogrammetric Engineering*, 130.

[Chen et al., 2018] Chen D, Li H, Xiao T, Yi S, Wang X (2018) Video Person Re-Identification With Competitive Snippet-Similarity Aggregation and Co-Attentive Snippet Embedding. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1169–1178.

[Cheng et al., 2016] Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1335–1344.

[Cheng & Cristani, 2014] Cheng DS, Cristani M (2014) Person Re-identification by Articulated Appearance Matching. In: *Person Re-Identification* (pp. 139–160).

[Chiang & Wang, 2014] Chiang A, Wang Y (2014) Human detection in fish-eye images using HOG-based detectors over rotated windows. In: *IEEE International Conference on Multimedia and Expo, Workshops (ICME-W)*: 1–6.

[Cho et al., 2017] Cho Y, Kim S, Park J, Lee K, Yoon K (2017) Joint Person Re-identification and Camera Network Topology Inference in Multiple Cameras. *Computing Research Repository (CoRR)*, abs/1710.00983.

[Cho & Yoon, 2016] Cho Y, Yoon K (2016) Improving Person Re-identification via Pose-Aware Multi-shot Matching. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1354–1362.

[Cho & Yoon, 2018] Cho Y, Yoon K (2018) PaMM: Pose-Aware Multi-Shot Matching for Improving Person Re-Identification. *IEEE Transactions on Image Processing (TIP)*, 27 (8): 3739–3752.

[Cordts et al., 2016] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 3213–3223.

[Cortes & Vapnik, 1995] Cortes C, Vapnik V (1995) Support-Vector Networks. *Machine Learning*, 20 (3): 273–297.

[Demiröz et al., 2012] Demiröz BE, Ari I, Eroglu O, Salah AA, Akarun L (2012) Feature-based tracking on a multi-omnidirectional camera dataset. In: *International Symposium on Communications, Control and Signal Processing (ISCCSP)*: 1–5.

[Deng et al., 2009] Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) ImageNet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 248–255.

[Denman et al., 2015] Denman S, Fookes C, Ryan D, Sridharan S (2015) Large scale monitoring of crowds and building utilisation: A new database and distributed approach. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*: 1–6.

[Eberle, 2018] Eberle A (2018) Pose-Driven Deep Models for Person Re-Identification. Master's thesis, Karlsruhe Institute of Technology, published on arxiv.

[Epanechnikov & Seckler, 1969] Epanechnikov VA, Seckler B (1969) Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability and its Applications*, 14: 153–158.

[Farenzena et al., 2010] Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2360–2367.

[Fergnani et al., 2016] Fergnani F, Alletto S, Serra G, Mira JD, Cucchiara R (2016) Body Part Based Re-Identification from an Egocentric Perspective. In: *IEEE Conference on Computer Vision and Pattern Recognition, Workshops (CVPR-W)*: 355–360.

[Förstner & Wrobel, 2016] Förstner W, Wrobel BP (2016) *Photogrammetric Computer Vision – Statistics, Geometry, Orientation and Reconstruction*. Springer.

[Freund & Schapire, 1997] Freund Y, Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 55 (1): 119–139.

[Frontex, 2011] Frontex (2011) Application of surveillance tools to Border Surveillance 'Concept of Operations'. `https://ec.europa.eu/research/participants/portal/doc/call/fp7/fp7-space-2012-1/31341-2011_concept_of_operations_for_the_common_application_of_surveillance_tools_in_the_context_of_eurosur_en.pdf`.

[García et al., 2015] García J, Martinel N, Micheloni C, Vicente AG (2015) Person Re-Identification Ranking Optimisation by Discriminant Context Information Analysis. In: *IEEE International Conference on Computer Vision (ICCV)*: 1305–1313.

[Geng et al., 2016] Geng M, Wang Y, Xiang T, Tian Y (2016) Deep Transfer Learning for Person Re-identification. *Computing Research Repository (CoRR)*, abs/1611.05244.

[Geyer & Daniilidis, 2000] Geyer C, Daniilidis K (2000) A Unifying Theory for Central Panoramic Systems and Practical Implications. In: Vernon D (ed) *European Conference on Computer Vision (ECCV)*: 445–461.

[Gheissari et al., 2006] Gheissari N, Sebastian TB, Hartley RI (2006) Person Reidentification Using Spatiotemporal Appearance. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1528–1535.

[Girshick et al., 2018] Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K (2018) Detectron. `https://github.com/facebookresearch/detectron`.

[Glorot et al., 2011] Glorot X, Bordes A, Bengio Y (2011) Deep Sparse Rectifier Neural Networks. In: *International Conference on Artificial Intelligence and Statistics AISTATS*: 315–323.

[Gong et al., 2014] Gong S, Cristani M, Yan S, Loy CC, eds (2014) *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition. Springer.

[Goodfellow et al., 2016] Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Goodfellow et al., 2014] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y (2014) Generative Adversarial Networks. *Computing Research Repository (CoRR)*.

[Gray et al., 2007] Gray D, Brennan S, Tao H (2007) Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In: *International Workshop on Performance Evaluation for Tracking and Surveillance*

[Gray & Tao, 2008] Gray D, Tao H (2008) Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: *European Conference on Computer Vision (ECCV)*: 262–275.

[Guo et al., 2016] Guo Y, Zhang L, Hu Y, He X, Gao J (2016) MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In: *European Conference on Computer Vision (ECCV)*: 87–102.

[Hakeem et al., 2012] Hakeem A, Gupta H, Kanaujia A, Choe TE, Gunda K, Scanlon A, Yu L, Zhang Z, Venetianer P, Rasheed Z, Haering N (2012) *Video Analytics for Business Intelligence*, (pp. 309–354). Springer.

[He et al., 2017] He K, Gkioxari G, Dollár P, Girshick RB (2017) Mask R-CNN. In: *IEEE International Conference on Computer Vision (ICCV)*: 2980–2988.

[He et al., 2016] He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 770–778.

[Hermans et al., 2017] Hermans A, Beyer L, Leibe B (2017) In Defense of the Triplet Loss for Person Re-Identification. *Computing Research Repository (CoRR)*, abs/1703.07737.

[Hochreiter & Schmidhuber, 1997] Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Computation*, 9 (8): 1735–1780.

[Huang & Russell, 1997] Huang T, Russell SJ (1997) Object Identification in a Bayesian Context. In: *International Joint Conference on Artificial Intelligence (IJCAI)*: 1276–1283.

[Ibrahim, 2011] Ibrahim E (2011) *Vision based tracking in team sports*. PhD thesis, University of Paderborn.

[IEC, 2014] IEC (2014) *IEC 62676-4:2014 Video surveillance systems for use in security applications - Part 5: Data specifications and image quality performance for camera devices*. International Electrotechnical Commission, Standard.

[Imani & Soltanizadeh, 2016] Imani Z, Soltanizadeh H (2016) Person Reidentification Using Local Pattern Descriptors and Anthropometric Measures From Videos of Kinect Sensor. *IEEE Sensors Journal*, 16 (16): 6227–6238.

[Imran, 2014] Imran A (2014) *Person detection using wide angle overhead cameras*. PhD thesis, University of Southampton.

[Ioffe & Szegedy, 2015] Ioffe S, Szegedy C (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *International Conference on Machine Learning (ICML)*: 448–456.

[Jović et al., 2006] Jović M, Hatakeyama Y, Dong F, Hirota K (2006) Image Retrieval Based on Similarity Score Fusion from Feature Similarity Ranking Lists. In: Wang L, Jiao L, Shi G, Li X, Liu J (eds) *Fuzzy Systems and Knowledge Discovery*: 461–470.

[Kang, 2000] Kang SB (2000) Catadioptric Self-Calibration. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 201–207.

[Kingma & Ba, 2015] Kingma DP, Ba J (2015) Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations (ICLR)*

[Kingslake, 1985] Kingslake R (1985) Development of the Photographic Objective. 0531.

[Kittler et al., 1998] Kittler J, Hatef M, Duin RPW, Matas J (1998) On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20 (3): 226–239.

[Knorr, 2018] Knorr M (2018) *Self-Calibration of Multi-Camera Systems for Vehicle Surround Sensing*. PhD thesis, Karlsruher Institut für Technologie (KIT).

[Köstinger et al., 2012] Köstinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large Scale Metric Learning from Equivalence Constraints. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2288–2295.

[Leal-Taixé et al., 2015] Leal-Taixé L, Milan A, Reid I, Roth S, Schindler K (2015) MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *Computing Research Repository (CoRR)*.

[Leng et al., 2015] Leng Q, Hu R, Liang C, Wang Y, Chen J (2015) Person re-identification with content and context re-ranking. *Multimedia Tools Appl.*, 74 (17): 6989–7014.

[Li et al., 2018] Li S, Bak S, Carr P, Wang X (2018) Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-Identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 369–378.

[Li et al., 2012] Li W, Zhao R, Wang X (2012) Human Reidentification with Transferred Metric Learning. In: *Asian Conference on Computer Vision (ACCV)*: 31–44.

[Li et al., 2014] Li W, Zhao R, Xiao T, Wang X (2014) DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 152–159.

[Liao et al., 2015] Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by Local Maximal Occurrence representation and metric learning. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2197–2206.

[Liao et al., 2010] Liao S, Zhao G, Kellokumpu V, Pietikainen M, Li SZ (2010) Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1301–1306.

[Liciotti et al., 2017] Liciotti D, Paolanti M, Frontoni E, Mancini A, Zingaretti P (2017) *Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration*, (pp. 1–11). Springer International Publishing: Cham.

[Lin et al., 2014] Lin T, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: Common Objects in Context. In: *European Conference on Computer Vision (ECCV)*: 740–755.

[Lovric et al., 2000] Lovric M, Min-Oo M, Ruh EA (2000) Multivariate Normal Distributions Parametrized as a Riemannian Symmetric Space. *Journal of Multivariate Analysis*, 74 (1): 36–48.

[Lucas & Kanade, 1981] Lucas BD, Kanade T (1981) An Iterative Image Registration Technique with an Application to Stereo Vision. In: *International Joint Conference on Artificial Intelligence, IJCAI*: 674–679.

[Luo et al., 2019] Luo H, Gu Y, Liao X, Lai S, Jiang W (2019) Bag of Tricks and A Strong Baseline for Deep Person Re-identification. *Computing Research Repository (CoRR)*, abs/1903.07071.

[Ma et al., 2016] Ma L, Liu H, Hu L, Wang C, Sun Q (2016) Orientation Driven Bag of Appearances for Person Re-identification. *Computing Research Repository (CoRR)*, abs/1605.02464.

[Matsukawa et al., 2016] Matsukawa T, Okabe T, Suzuki E, Sato Y (2016) Hierarchical Gaussian Descriptor for Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1363–1372.

[Matsukawa et al., 2019] Matsukawa T, Okabe T, Suzuki E, Sato Y (2019) Hierarchical Gaussian Descriptors with Application to Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[McLaughlin et al., 2016] McLaughlin N, del Rincón JM, Miller PC (2016) Recurrent Convolutional Network for Video-Based Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1325–1334.

[Mei, 2007] Mei C (2007) *Couplage Vision Omnidirectionnelle et Télémétrie Laser pour la Navigation en Robotique/Laser-Augmented Omnidirectional Vision for 3D Localisation and Mapping*. PhD thesis, INRIA Sophia Antipolis, Project-team ARobAS.

[Mei & Rives, 2007] Mei C, Rives P (2007) Single View Point Omnidirectional Camera Calibration from Planar Grids. In: *IEEE International Conference on Robotics and Automation (ICRA)*: 3945–3950.

[Mignon & Jurie, 2012] Mignon A, Jurie F (2012) PCCA: A new approach for distance learning from sparse pairwise constraints. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2666–2672.

[Milan et al., 2016] Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K (2016) MOT16: A Benchmark for Multi-Object Tracking. *Computing Research Repository (CoRR)*. arXiv: 1603.00831.

[Moghaddam et al., 2000] Moghaddam B, Jebara T, Pentland A (2000) Bayesian face recognition. *Pattern Recognition*, 33 (11): 1771–1782.

[Nambiar, 2017] Nambiar AM (2017) *Towards automatic long term Person Re-identification System in video surveillance*. PhD thesis, Universidade de Lisboa.

[Neyman & Pearson, 1928] Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Part 2. *Biometrika*, 20A (1-2): 175–240.

[Pedagadi et al., 2013] Pedagadi S, Orwell J, Velastin SA, Boghossian BA (2013) Local Fisher Discriminant Analysis for Pedestrian Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 3318–3325.

[Plantinga, 1961] Plantinga A (1961) Things and Persons. *The Review of Metaphysics*, 14 (3): 493–519.

[Ristani et al., 2016] Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In: *European Conference on Computer Vision, Workshops (ECCV-W)*: 17–35.

[Saal et al., 1980] Saal F, Downey R, Lahey M (1980) Rating the Ratings: Assessing the Psychometric Quality of Rating Data. *Psychological Bulletin*, 88: 413–428.

[Salam, 2019] Salam RT (2019) *Multi-View Person Re-Identification*. Master's Thesis, supervised by Bosch company, University of Dortmund (unpublished).

[Sarfraz et al., 2018] Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R (2018) A Pose-Sensitive Embedding for Person Re-Identification With Expanded Cross Neighborhood Re-Ranking. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 420–429.

[Scaramuzza, 2014] Scaramuzza D (2014) Omnidirectional Camera. In: Katsushi I (ed) *Computer Vision: A Reference Guide*: 552–560.

[Schön et al., 2018] Schön S, Brenner C, Alkhatib H, Coenen M, Dbouk H, Garcia-Fernandez N, Fischer C, Heipke C, Lohmann K, Neumann I, Nguyen U, Paffenholz JA, Peters T, Rottensteiner F, Schachtschneider J, Sester M, Sun L, Vogel S, Voges R, Wagner B (2018) Integrity and Collaboration in Dynamic Sensor Networks. *Sensors*, 18 (7).

[Schönbein, 2014] Schönbein M (2014) *Omnidirectional Stereo Vision for Autonomous Vehicles*. PhD thesis, Karlsruhe Institute für Technology (KIT).

[Schroff et al., 2015] Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: A unified embedding for face recognition and clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 815–823.

[Simonyan & Zisserman, 2014] Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computing Research Repository (CoRR)*, abs/1409.1556.

[Song et al., 2018] Song C, Huang Y, Ouyang W, Wang L (2018) Mask-Guided Contrastive Attention Model for Person Re-Identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1179–1188.

[STANAG, 1998] STANAG (1998) Nato Standardization Agreement 3769; Minimum resolved object sizes and scales for imagery interpretation.

[Statista.com, 2018] Statista.com (2018) Security & surveillance technology. `https://www.statista.com/statistics/879212/china-market-size-of-video-surveillance/`.

[Strauß et al., 2014] Strauß T, Ziegler J, Beck J (2014) Calibrating multiple cameras with non-overlapping views using coded checkerboard targets. In: *International IEEE Conference on Intelligent Transportation Systems (ITSC)*: 2623–2628.

[Su et al., 2017] Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-Driven Deep Convolutional Model for Person Re-identification. In: *IEEE International Conference on Computer Vision (ICCV)*: 3980–3989.

[Sun & Zheng, 2019] Sun X, Zheng L (2019) Dissecting Person Re-Identification From the Viewpoint of Viewpoint. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 608–617.

[Svoboda & Pajdla, 2002] Svoboda T, Pajdla T (2002) Epipolar Geometry for Central Catadioptric Cameras. *International Journal of Computer Vision (IJCV)*, 49 (1): 23–37.

[Szegedy et al., 2016] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 2818–2826.

[Szeliski, 2011] Szeliski R (2011) *Computer Vision - Algorithms and Applications*. Texts in Computer Science. Springer.

[Telegraph, 2013] Telegraph T (2013) One surveillance camera for every 11 people in Britain, says CCTV survey. `https://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html`.

[Thibault, 2010] Thibault S (2010) *Panomoprh Based Panoramic Vision Sensors*. Sciyo.

[Tian et al., 2018] Tian M, Yi S, Li H, Li S, Zhang X, Shi J, Yan J, Wang X (2018) Eliminating Background-Bias for Robust Person Re-Identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 5794–5803.

[van de Weijer et al., 2009] van de Weijer J, Schmid C, Verbeek JJ, Larlus D (2009) Learning Color Names for Real-World Applications. *IEEE Transactions on Image Processing (TIP)*, 18 (7): 1512–1523.

[Varior et al., 2016] Varior RR, Haloi M, Wang G (2016) Gated Siamese Convolutional Neural Network Architecture for Human Re-identification. In: *European Conference on Computer Vision (ECCV)*: 791–808.

[Weinberger & Saul, 2009] Weinberger KQ, Saul LK (2009) Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.*, 10: 207–244.

[Xiao et al., 2016] Xiao T, Li H, Ouyang W, Wang X (2016) Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1249–1258.

[Yang et al., 2014] Yang Y, Yang J, Yan J, Liao S, Yi D, Li SZ (2014) Salient Color Names for Person Re-identification. In: *European Conference on Computer Vision (ECCV)*: 536–551.

[Yi et al., 2014] Yi D, Lei Z, Liao S, Li SZ (2014) Deep Metric Learning for Person Re-identification. In: *International Conference on Pattern Recognition (ICPR)*: 34–39.

[Ying & Hu, 2004] Ying X, Hu Z (2004) Can We Consider Central Catadioptric Cameras and Fisheye Cameras within a Unified Imaging Model. In: *European Conference on Computer Vision (ECCV)*: 442–455.

[Zajdel et al., 2005] Zajdel W, Zivkovic Z, Kröse BJA (2005) Keeping Track of Humans: Have I Seen This Person Before? In: *IEEE International Conference on Robotics and Automation (ICRA)*: 2081–2086.

[Zhang et al., 2016] Zhang L, Xiang T, Gong S (2016) Learning a Discriminative Null Space for Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1239–1248.

[Zhao et al., 2013] Zhao R, Ouyang W, Wang X (2013) Unsupervised Salience Learning for Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 3586–3593.

[Zheng et al., 2016a] Zheng L, Bie Z, Sun Y, Wang J, Su C, Wang S, Tian Q (2016a) MARS: A Video Benchmark for Large-Scale Person Re-identification. In: *European Conference on Computer Vision (ECCV)*: 868–884.

[Zheng et al., 2015a] Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015a) Scalable Person Re-identification: A Benchmark. In: *IEEE International Conference on Computer Vision (ICCV)*: 1116–1124.

[Zheng et al., 2015b] Zheng L, Wang S, Tian L, He F, Liu Z, Tian Q (2015b) Query-Adaptive Late Fusion for Image Search and Person Re-identification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 1741–1750.

[Zheng et al., 2016b] Zheng L, Yang Y, Hauptmann AG (2016b) Person Re-identification: Past, Present and Future. *Computing Research Repository (CoRR)*, abs/1610.02984.