Robuste Detektion, Verfolgung und Wiedererkennung von Personen in Videodaten mit niedriger Auflösung

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Dipl.-Inform. Jürgen Metzler

Tag der mündlichen Prüfung:	29.06.2018
Erster Gutachter:	Prof. Dr.-Ing. Jürgen Beyerer
Zweiter Gutachter:	Prof. Dr.-Ing. Rainer Stiefelhagen

Zusammenfassung

Im Bereich der Videoüberwachung bietet die steigende Zahl an Videokameras ein enormes Potential aus Sicht der zivilen Sicherheit. Mit der zunehmenden Menge an Bilddaten in diesem Bereich wächst auch die Chance, Straftaten besser aufklären oder sogar vermeiden zu können. Allerdings ist dafür ein immenser Aufwand für die Auswertung der Bilder erforderlich, die oft nicht mehr vollständig ohne Computerunterstützung durch Personen gesichtet werden können. Folglich wächst auch der Bedarf an Verfahren für die (semi-)automatische und intelligente Bildauswertung, wobei die Detektion, Verfolgung und Wiedererkennung von Personen Aufgabenschwerpunkte bilden.

Diese Bildauswerteaufgaben stellen für die Prävention und Aufklärung von Straftaten eine sehr hilfreiche Unterstützung, aber auch schwierige Herausforderung dar. Insbesondere die Aufklärung in Multi-Kamera-Netzwerken ist eine schwierige Aufgabe, da die Personen in den einzelnen Kameras rasch und zuverlässig wiedererkannt werden müssen, um Tathergänge vollumfänglich rekonstruieren zu können. Erschwerend kommt für die Bildauswerteaufgaben hinzu, dass in der Videoüberwachung oft nur niedrig aufgelöste Bilddaten zur Verfügung stehen, aus denen beispielsweise keine biometrischen Merkmale von Personen extrahiert werden können. Darüber hinaus gibt es noch zahlreiche weitere Herausforderungen wie z.B. unterschiedliche Kameraausrichtungen, -typen und -konfigurationen sowie Unterschiede in der Umgebungsbeleuchtung.

Diese Arbeit umfasst Methoden und Verbesserungen auf Basis neuartiger Personenrepräsentationen für die Detektion, Verfolgung und erscheinungs-

basierte Wiedererkennung von Personen, die typische Schwierigkeiten, die sich im Bereich der Videoüberwachung stellen, berücksichtigen und insbesondere für die Anwendung auf niedrig aufgelösten Bildern geeignet sind. Die Ansätze basieren auf einem einheitlichen Bildauswerte-Rahmenwerk, das auf Kovarianzdeskriptoren als Einzelbild-Deskriptoren und einer riemannschen Mannigfaltigkeit von symmetrischen, positiv definiten Matrizen aufbaut. Der große Vorteil eines auf Kovarianzdeskriptoren basiertes Rahmenwerks ist seine Offenheit für beliebige Bildmerkmale. Es hat sich im Rahmen dieser Arbeit für die Personenrepräsentation in niedrig aufgelösten Bildern sowie gegenüber anderen Herausforderungen als geeignet erwiesen. Die Verfahren wurden in drei unterschiedlichen Anwendungsdomänen erarbeitet, um die Generalisierungsfähigkeit des zugrunde liegenden Bildauswerte-Rahmenwerks aufzuzeigen. Die betrachteten Datensätze bestehen aus Wärme- und Farbbildern, die sowohl mittels stationären als auch mobilen Boden- und Luftkameras akquiriert wurden.

Ziel der **Personendetektion** ist es, Personen in Einzelbildern anhand ihrer Körperteile zu detektieren. Um einer niedrigen Auflösung und einer minderen Datenqualität gerecht zu werden, werden Körperteile durch Kovarianzdeskriptoren repräsentiert. Zudem werden die Kovarianzdeskriptoren einer Körperteilklasse als Mannigfaltigkeit aufgefasst, um die Robustheit des Körperteildetektors zu erhöhen. Der Kern des Verfahrens ist ein Manifold Learning Algorithmus, welcher die einzelnen Körperteilmannigfaltigkeiten überwacht lernt, wobei eine riemannsche Metrik des riemannschen Raums der symmetrischen, positiv definiten Kovarianzmatrizen zugrunde gelegt wird. Dabei wird das Ziel verfolgt, die Intra-Diskriminanz von Bildern einer Körperteilklasse zu senken und die Inter-Diskriminanz zwischen unterschiedlichen Körperteilmannigfaltigkeiten zu erhöhen.

Die Arbeiten im Bereich des **Videotrackings** umfassen Verbesserungen eines auf Kovarianzdeskriptoren basiertes Trackingverfahren, mit dem Ziel, einzelne Personen robust in Menschenmengen verfolgen zu können. Das verbesserte Verfahren repräsentiert einzelne Personen durch jeweils einen Mittelwert-Kovarianzdeskriptor, der sich aus mehreren einzelnen Kovarianzdeskriptoren bestimmt. Die Auswahl der einzelnen Kovarianzdeskriptoren erfolgt anhand statistischer Eigenschaften hinsichtlich der Mannigfaltigkeit der Deskriptoren, um eine — sich über die Zeit anpassende — robuste Personenrepräsentation zu erreichen. Zudem wird eine Mahalanobis-Distanz für die Zuordnung der Personen von Bild-zu-Bild

verwendet, die im Raum der positiv definiten Kovarianzmatrizen definiert ist.

Den Schwerpunkt der Arbeit bildet die **erscheinungsbasierte Personenwiedererkennung**. Die erscheinungsbasierte Personenwiedererkennung umfasst eine bildsequenzbasierte Personenrepräsentation für eine effiziente Überprüfung, ob mehrere Ganzkörpersequenzen von derselben Person stammen, um Personen kameraübergreifend wiedererkennen oder in Bilddatensätzen suchen zu können. Darin unterscheidet sich dieser Ansatz von den meisten existierenden erscheinungsbasierten Personenwiedererkennungsverfahren, die ausschließlich Einzelbilder verwenden. Die bildsequenzbasierte Personenrepräsentation bestimmt sich aus mehreren Kovarianzdeskriptoren, wodurch sowohl verschiedene Ausprägungen der Personenerscheinung als auch Bildausschnitte niedriger Qualität kompensiert werden können. Um den Herausforderungen weiter zu begegnen, wird eine Manifold Learning basierte Neusortierung von Teilergebnissen untersucht. Dazu wird unter der Annahme, dass Deskriptoren einer Person auf einer Mannigfaltigkeit liegen, eine einzelbildbasierte Neusortierung durchgeführt.

In dieser Arbeit wird im Vergleich zu vielen anderen Ansätzen eine unüberwachte Personenwiedererkennung verfolgt, die ausschließlich die Bilddaten benötigt, die für die Aufklärung relevant sind. Für den Fall existenter externer Daten werden abschließend in dieser Arbeit überwacht gelernte Möglichkeiten vorgestellt, um eine verbesserte kovarianzdeskriptorbasierte Repräsentation zu trainieren. Dazu werden künstliche faltende neuronale Netze verwendet, die logarithmierte Kovarianzdeskriptoren lernen oder handentworfene Kovarianzdeskriptoren im Lernprozess mit berücksichtigen, um die Stärken der Kovarianzdeskriptoren mit gelernten Merkmalen zu verknüpfen.

Die Auswertung der vorgestellten Detektions- und Trackingverfahren erfolgt auf selbst akquirierten Bilddaten relevanter Anwendungs- und Überwachungsszenarien. Zur Evaluation der erscheinungsbasierten Personenwiedererkennungsverfahren werden sowohl selbst akquirierte als auch öffentliche Datensätze verwendet. In allen drei Bildauswertebereichen konnten mit den neuartigen Personenrepräsentationen gegenüber relevanten Referenzverfahren bessere Ergebnisse erzielt werden.

Abstract

In the field of video surveillance, the increasing number of video cameras offers enormous potential from the point of view of civil security. With the increasing amount of image data, the chance to better prevent and investigate crimes also increases. However, this requires big effort for the exploitation of the images, which is prohibitive without computer support by humans. Consequently, the need for methods for (semi-)automatic and intelligent image analysis is growing in video surveillance, where the detection, tracking and re-identification of people are the main focuses.

These image exploitation tasks are very helpful for the prevention and investigation of criminal offenses, but they are also challenging tasks. In particular, the investigation in multi-camera networks is a major challenge, since the people in the individual cameras must be quickly and reliably re-identified in order to fully reconstruct the circumstances of a crime. This is aggravated by the fact that in the field of video surveillance often only low-resolution image data are available, which is a major challenge for the image exploitation tasks. For instance, no biometric features of people can be extracted from the images. In addition, there are many other challenges such as different camera orientations, types and configurations as well as differences in ambient lighting, which must be taken into account when developing image exploitation methods. The quality of image data — even with high-resolution cameras — is generally not the same as the quality of e.g. studio shots or professional shots for the media, which are in contrast to the video surveillance cooperative scenarios with controlled lighting, focused persons etc.

The subjects of this thesis are novel methods and enhancements for the detection, tracking and appearance-based re-identification of persons that consider typical challenges in the field of video surveillance, particularly for low-resolution images. The methods are built upon a unified framework based on covariance descriptors as image descriptors and a Riemannian manifold of symmetric, positive definite matrices. The big advantage of a framework based on covariance descriptors is its compatibility to any image features. It has been shown to be suitable for low-resolution person representation and other challenges in this work. The methods were developed in three different application domains to show the generalizability of the underlying framework, where the data sets consist of thermal and color images acquired by both stationary and mobile ground and aerial cameras.

The aim of the person detection is to detect persons in single images on the basis of their body parts, which are represented by covariance descriptors to handle low resolution and degraded data quality. In addition, the covariance descriptors of a body subclass are considered as a manifold in order to increase the robustness of the image-based body detector. The core of the method is a Manifold Learning algorithm, which learns the individual body part manifolds in a supervised manner, using a Riemannian metric of the Riemannian space of the symmetric, positive definite matrices. The aim is to decrease the intra-discriminance of images of a body part class and to increase the inter-discriminance between different body part manifolds.

The work in the field of image-based tracking includes improvements to a covariance descriptor-based tracking method, with the goal of robustly tracking individuals in crowds. The improved tracking method represents individual persons by means of a mean covariance descriptor determined from individual covariance descriptors. The selection of the individual covariance descriptors is done on the basis of statistical properties regarding the multiplicity of descriptors in order to achieve a robust person representation that adapts over time. In addition, Mahalanobis distance is used for the assignment of the persons from image to image, which is defined in the space of positive definite covariance matrices.

The main focus of the work is the appearance-based person re-identification. Appearance-based person re-identification includes an image sequence based person representation for an efficient verification of whether multiple

whole body sequences originate from the same person in order to re-identify persons across cameras or to search persons in image data sets. Therein, this approach differs from most existing appearance-based person re-identification methods that use only single images. The image sequence-based person representation is determined by several covariance descriptors in order to handle different appearances of the person as well as to compensate for image regions of poor image quality. To further address the challenges, a Manifold Learning based re-ranking of the partial results is investigated. For this purpose, an image-based re-ranking is carried out under the assumption that individual person images lie on a manifold.

In this work, in comparison to many other approaches, an unsupervised person re-identifiaction method that only requires the image data relevant for the crime investigation is proposed. In the case of existing external data, supervised learning opportunities to train an improved covariance descriptor-based representation are finally presented in this work. Deep convolutional neural networks are used that learn the logarithmics of covariance descriptors or consider hand-crafted covariance descriptors in the learning process in order to combine the strengths of the covariance descriptors with learned features.

The evaluations of the proposed detection and tracking methods are conducted on self-acquired image data of relevant applications and surveillance scenarios. Both self-acquired and public data sets are used to evaluate the appearance-based person re-identification approaches. The developed methods have been shown to achieve better results than relevant reference methods.

Inhaltsverzeichnis

1

Einleitung

Im Bereich der zivilen Sicherheit spielen Kameras eine immer wichtigere Rolle. Allein in Deutschland ist nach Recherchen und Schätzungen von *IHS Markit*[1] im Jahr 2016 die Zahl der Kameras um 800.000 gestiegen, was einer Steigerung von 20% entspricht. Eine andere Recherche von *Frost & Sullivan* schätzt, dass der US-Videoüberwachungsmarkt bis 2023 um jährlich 5% wachsen wird [Glo17]. Der prozentuale Anteil der Kameras in diesem US-Markt lag 2016 nach dieser Recherche bei 48%. Mit der steigenden Zahl der Kameras wächst auch die Videodatenmenge, die gesichtet werden muss.

1.1 Motivation

Die heutzutage anfallenden Bilddaten im Bereich der Videoüberwachung können oft nicht mehr vollständig durch Personen ohne Computerunterstützung oder nur durch erheblichen personellen Aufwand gesichtet werden [Qu18]. Intelligente Bildauswerteverfahren zur (semi-)automatischen Auswertung oder — aus Sicht der Datensparsamkeit — Vorfilterung der Bilder können hierbei unterstützen oder sogar selbstständig Aufgaben lösen. Die

[1] `https://technology.ihs.com/583518/video-surveillance-cameras-installed-in-germany-to-increase-20-percent-this-year`

Einsatzbereiche der automatischen Bildauswerteverfahren in der Videoüberwachung sind dabei vielfältig: Sie reichen z.B. von einer einfachen Bewegungsdetektion in Bildern für eine eigenständige Alarmgenerierung bei Bewegungen, über die selbständige Erkennung auffälliger Personenbewegungen, was eine Personendetektion und Klassifizierung der Bewegung erfordert, bis hin zu automatischen komplexen Analysen von Personengangarten, um z.B. auf die Identitäten der Personen schließen zu können. Dies sind nur drei Beispiele von zahlreichen weiteren Aufgaben im Bereich der intelligenten Videoüberwachung.

Viele bildbasierte Videoüberwachungslösungen haben eines gemein: das Interesse an Personen im Bild. Selbst wenn die Aufgabe die Detektion herrenloser Gepäckstücke ist, möchte man in sicherheitsrelevanten Szenarien einen automatisch generierten Hinweis erhalten, wer dieses Gepäckstück wann abgestellt hat. Diese Dissertationsschrift beschäftigt sich mit der automatischen bildbasierten Detektion und Verfolgung von Personen sowie der erscheinungsbasierten Wiedererkennung von bereits bekannten Personen anhand von Ganzkörpermerkmalen wie z.B. Kleiderfarbe. Aufgrund der Relevanz von Personen in sicherheitsrelevanten Aufgabenstellungen sind das sehr hilfreiche, allerdings auch schwierige Aufgaben. Je nach Anwendungsfall, z.B. wenn es um die automatische Auswertung von Personenbildern und -handlungen im Zusammenhang mit Straftaten geht, sind robuste Verfahren zwingend notwendig. Für einen praktischen Nutzen sind hohe Genauigkeiten bei personenbezogenen Bildauswerteverfahren nötig. Bei dem erwähnten Anwendungsfall der Detektion von herrenlosen Gepäckstücken könnte beispielsweise eine zusätzliche automatisierte Suche nach weiteren Vorkommen der gesuchten Person erfolgen. Damit hätte man die Aufgabe der Detektion eines herrenlosen Gepäckstücks einschließlich der Feststellung möglicher Aufenthaltsorte der Person, die das Gepäckstück abgestellt hat, computergestützt gelöst. Die Wahrscheinlichkeit einer Verwechslung der Person sollte dabei allerdings minimiert sein.

Bei der Erarbeitung robuster Bildauswerteverfahren müssen zahlreiche Herausforderungen gemeistert werden. Im folgenden Abschnitt werden zunächst die drei Bildauswerteaufgaben kurz vorgestellt, die im Rahmen dieser Dissertationsschrift betrachtet werden. Der darauffolgende Abschnitt 1.3 gibt dann einen kurzen Überblick der Herausforderungen, die sich bei diesen Aufgaben stellen.

1.2 Bildauswerteaufgaben

Diese Dissertationsschrift behandelt drei unterschiedlichen Anwendungen im Videoüberwachungsbereich. Die automatischen und bildbasierten Ansätze, die in diesem Rahmen realisiert wurden, sind

1. die Personendetektion anhand von Körperteilen,
2. die Verfolgung bzw. das Tracking einzelner Personen in Menschenmengen und
3. die erscheinungsbasierte Personenwiedererkennung.

Die Verfahren wurden in drei unterschiedlichen Anwendungsdomänen erarbeitet, um die Generalisierungsfähigkeit des zugrunde liegenden Bildauswerte-Rahmenwerks aufzuzeigen. Die betrachteten Datensätze decken dabei sowohl Ferninfrarot- als auch visuell-optische Farbbilder von sowohl stationären als auch mobilen Boden- und Luftkameras ab. Abbildung 1.1 zeigt repräsentative Beispielbilder aus den verwendeten Datensätzen.

Personendetektion. Die Aufgabenstellung ist die einzelbildbasierte Detektion von Personen in Wärmebildern anhand des Kopfs, Torsos und/oder der Beine. Das Anwendungsszenario ist im Kontext der Fahrzeugumfelderfassung die frühzeitige Detektion aller Personen in sicherheitsrelevanten Szenarien, wie beispielsweise Aufklärungsfahrten in Krisengebieten. Dabei sollen auch Personen detektiert werden, bei denen beispielsweise nur der Kopf im Kamerabild sichtbar ist (siehe z.B. mittleres Bild in der 1. Reihe in Abbildung 1.1).

Personentracking. Das Ziel im Rahmen des Personentrackings ist ein robustes Verfahren zu erarbeiten, das zuverlässig einzelne manuell markierte Personen in Luftbildern von Menschenmengen verfolgen kann. Die zweite Reihe in Abbildung 1.1 zeigt anhand von zwei Bildbeispielen typische Szenarien von Menschenmengen, die für diese Arbeit relevant sind. Das Verfahren soll zu Deeskalationszwecken eingesetzt werden sowie die Ergreifung und Beobachtung von Tatverdächtigen in Menschenmengen unterstützen. Bei unfriedlichem Verhalten in Menschenansammlungen geht

Abbildung 1.1: Beispielbilder aus den Datensätzen, die im Rahmen dieser Arbeit verwendet wurden. In der ersten Reihe sind exemplarisch drei Ferninfrarotbilder (Wärmebilder) aus dem Datensatz dargestellt, der im Rahmen der Körperteildetektion betrachtet wird. Die mittlere Reihe zeigt zwei repräsentative Luftbildaufnahmen von Menschenmengen des Tracking-relevanten Datensatzes. Die unterste Reihe zeigt typische Bildausschnitte von Personen, die bei der erscheinungsbasierten Wiedererkennung verarbeitet werden. Die Beispielbildausschnitte stammen aus sechs unterschiedlichen Datensätzen.

dieses Verhalten oft von wenigen einzelnen Personen (Störern) aus, die von der Menschenmenge getrennt werden sollten. Sowohl bei diesem Fall als auch bei der Ergreifung von Tatverdächtigen in Menschenmengen muss dafür ein günstiger Moment abgewartet werden. Dazu wird der Tatverdächtige in der Regel durchgängig manuell beobachtet. Dieser Prozess soll durch das Trackingverfahren unterstützt werden, wobei insbesondere wichtig ist, die Anzahl der Verwechslungen gering zu halten.

Personenwiedererkennung. Die Motivation hinter der Erforschung von Personenwiedererkennungsverfahren sind zwei Anwendungsfälle. Erstes Ziel ist die Wiedererkennung von Personen in Kameranetzwerken, d.h. beispielsweise die Wiedererkennung von Personen zwischen unterschiedlichen Kameras, deren Sichtbereiche sich nicht überlappen. Das wäre eine Wiedererkennung, die in (nahezu) Echtzeit funktionieren muss. Damit können beispielsweise einzelne Personen kameraübergreifend in Kameranetzwerken getrackt werden. Die zweite Anwendung ist die Suche von Personen in aufgezeichnetem Bildmaterial. In beiden Anwendungen wird eine erscheinungsbasierte Personenwiedererkennung verfolgt, d.h. eine Wiedererkennung beispielsweise anhand der Kleiderfarbe oder Muster auf der Kleidung. Die erscheinungsbasierte Personenwiedererkennung spielt im Bereich der Videoüberwachung eine große Rolle, da die Bilder trotz moderner Videokameras sehr oft keine biometrische Personenwiedererkennung zulassen, sei es durch eine zu geringe Auflösung, Bewegungsunschärfe oder aufgrund nicht sichtbarer Gesichter (siehe Reihe 3 in Abbildung 1.1). Im folgenden Abschnitt werden solche Herausforderungen näher erläutert.

1.3 Herausforderungen

Bildauswerteaufgaben für den Videoüberwachungsbereich, bei denen *die Person im Kamerabild* im Fokus steht, sind schwierige Aufgaben. Neben — trotz hochauflösender Kameras — sehr oft niedrigen Auflösung von Personen in den Bildern, gibt es noch zahlreiche weitere Herausforderungen, die bei der Erarbeitung von Verfahren für die Detektion, Verfolgung und Wiedererkennung von Personen berücksichtigt werden müssen. Die wichtigsten Herausforderungen werden im Folgenden kurz erläutert, wobei

sie in bildqualitäts- und anwendungsbezogene Herausforderungen unterteilt werden.

Herausforderungen bezüglich niedriger Bildqualität:

- **Niedrige Auflösung** von Personen resultiert aus dem Einsatz kostengünstiger Kameras. Aber auch hochauflösende Kameras werden in der Videoüberwachungsaufgabe so kosteneffizient installiert, dass sie möglichst viel von der Umgebung erfassen, was wiederum zu dem Ergebnis einer niedrigen Auflösung führt. Zudem sind oft — auch als Folge des gerade erwähnten Interesses — die Abstände der Kameras zu den Personen sehr groß. Dazu kommen noch Luftbildaufnahmen, die beispielsweise mit Hilfe von Quadrokoptern kostengünstig erzeugt sowie in Echtzeit übertragen werden können und somit immer interessanter für den Videoüberwachungsbereich werden. Dabei sind aber selbst bei hohen Bildauflösungen die Personen in der Regel niedrig aufgelöst. Die Abbildung 1.2 zeigt exemplarisch eine typische Aufnahme, die aus einer Höhe von ca. 25 Metern aufgenommen wurde. In Abbildung 1.3 sind zudem weitere Beispielbilder für typische Auflösungen in Videoüberwachungsszenarien dargestellt. Die Auflösungen sind oft so niedrig, dass keine biometrischen Merkmale von Personen aus den Bildern extrahiert werden können.

- **Unschärfe** kann z.B. durch eine hohe Belichtungszeit entstehen, wodurch schnelle Personenbewegungen unscharf abgebildet werden (Bewegungsunschärfe). Die Belichtungszeiten müssen bei Kameras in Innenräumen aufgrund schlechter Beleuchtungsbedingungen notwendigerweise oft hoch eingestellt werden. Eine andere häufige Ursache für Unschärfe ist die feste Fokussierung (auf beispielsweise den Hintergrund), so dass die Personen außerhalb des Fokus liegen können.

- **Schlechte Beleuchtungsbedingungen**, wie z.B. schlechte Ausleuchtungen in Innenräumen, können eine Bewegungsunschärfe verursachen. Zudem können sie Bildrauschen verursachen, insbesondere in niedrig aufgelösten Videos. Aber auch direkte Sonneneinstrahlung, die beispielsweise zu einer kompletten Überbelichtung der Szene führen kann, stellt eine Herausforderung dar. Eine starke Sonneneinstrahlung kann auch große Schatten erzeugen, die insbesondere

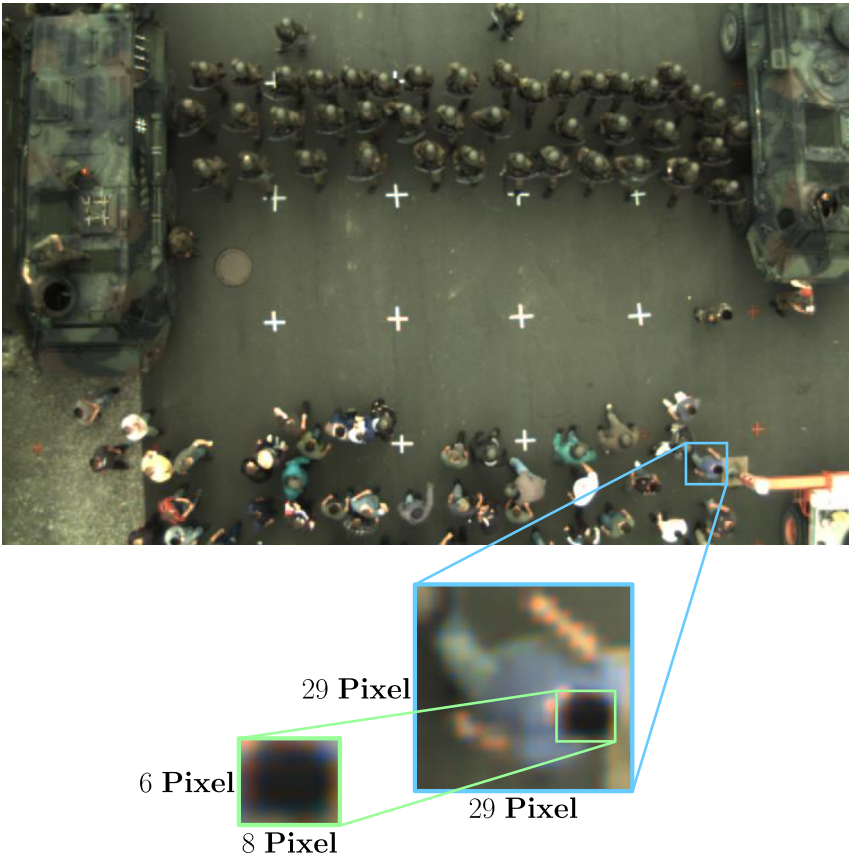


Abbildung 1.2: Bild aus dem Datensatz, der im Rahmen der Aufgabe *Personentracking in Menschenmengen* in Kapitel 6 verwendet wird. Eine Bildregion, in der eine Person abgebildet ist, ist exemplarisch vergrößert dargestellt. Der größere blau umrandete Bildausschnitt ist die Vergrößerung der Person und der hellgrün umrandete Ausschnitt links unten zeigt eine nochmalige Vergrößerung des Kopfs mit den jeweiligen Auflösungen im Originalbild.

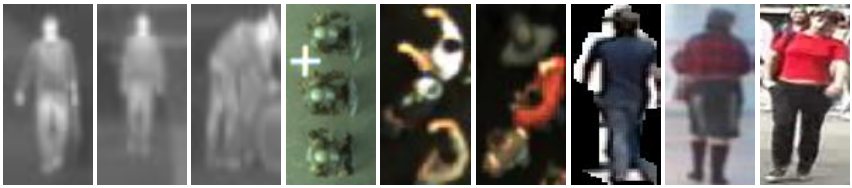


Abbildung 1.3: Die Bildausschnitte zeigen Personen in einer Auflösung, die typisch für Videoüberwachungsszenarien sind. Die Bildausschnitte sind der Anwendung entsprechend gruppiert: die linken drei sind für die Detektion, die mittleren drei für das Tracking und die rechten drei für die Personenwiedererkennung relevant. Die Bildausschnitte zeigen niedrig aufgelöste Personen, die den Schwerpunkt in dieser Arbeit bilden.

in Luftbildaufnahmen oft größere Bereiche im Bild belegen als die Personen selbst. Falls zudem der Kontrast der Personen zum Hintergrund schwach ist, haben gradientenbasierte Bildauswerteverfahren in der Regel sehr große Schwierigkeiten gute Ergebnisse zu erzielen.

- **Kompression** kann z.B. Farbinformationen verfälschen, die insbesondere bei kameraübergreifenden Aufgaben wie der erscheinungsbasierten Personenwiedererkennung zwischen verschiedenen Kameras eine sehr große Herausforderung darstellen kann. Unterschiedliche Kompressionseinstellungen zwischen den Kameras haben einen Einfluss auf beispielsweise die Segmentierung bewegter Objekte in Bildern (siehe Abbildung 3.4 in Abschnitt 3.3), was wiederum die Wiedererkennungsleistung beeinflusst. Gängige Kompressionsverfahren für Videoüberwachungskameras sind JPEG und h264, die zwar für das menschliche Auge in der Regel keine offensichtliche Qualitätseinbußen verursachen, bei den bildbasierten Verfahren, die auf den absoluten Pixelwerten aufbauen, jedoch zu einer deutlichen Verschlechterung des Auswerteergebnisses führen können.

- **Bildrauschen** entsteht durch stochastische Abweichungen der Pixelwerte aufgrund von Sensorrauschen. Schlechte Beleuchtungsbedingungen begünstigen das Bildrauschen, das insbesondere — ähnlich wie die Kompression — die erscheinungsbasierte Repräsentation von Personen in Bildern so stark beeinflussen kann, dass eine erschei-

nungsbasierte Wiedererkennung zwischen verschiedenen Kameras unmöglich wird.

- **Zeilensprünge** sind die Folge einer Reduktion der vertikalen Auflösung. Mittels Zeilensprungverfahren wird bei der Aufnahme jede zweite Bildzeile verworfen, abwechselnd beginnend mit der ersten oder zweiten Bildzeile. Anschließend werden zwei aufeinanderfolgende Bilder wieder zu einem Bild kombiniert, um eine Verringerung des Bildflimmerns für den menschlichen Betrachter zu erreichen. Aus Sicht der Bildauswertung bedeutet das allerdings einen Informationsverlust.

- **Linsenverzerrungen** treten oft bei preiswerten Kameraobjektiven auf, insbesondere bei Weitwinkelobjektiven, und können beispielsweise die Personenkontur so verändern, dass die Detektionsleistung sinkt.

- **Farbsäume** treten z.B. bei Einchip-Farbkameras auf, bei denen die Rot-, Grün- und Blauwerte der Bildpunkte aus den umgebenden Pixeln interpoliert werden, die jeweils nur eine Farbinformation abbilden können. Solche Kameras sind in der Regel kompakter und günstiger und sehr weit verbreitet. Die chromatische Aberration bei Linsen kann eine weitere Ursache für Farbsäume sein.

Anwendungsbezogene Herausforderungen:

- **Unterschiedliche Kameraperspektiven** oder verschiedene Ansichten und Posen von Personen können ein Problem darstellen, wenn die Bildauswerteverfahren beispielsweise auf Bildern entwickelt und getestet wurden, die keine ähnliche Perspektiven, Ansichten oder Posen der Personen aufzeigen.

- **Unterschiedliche Kameratypen und -konfigurationen** stellen für das Multi-Kamera-Tracking und für die erscheinungsbasierte Personenwiedererkennung zwischen unterschiedlichen Kameras eine große Herausforderung dar. Unterschiedliche Kameratypen und -konfigurationen können dieselbe Person beispielsweise farblich stark unterschiedlich abbilden, was eine deutliche Senkung der Wiedererkennungsleistung zur Folge hat.

- **Niedrige Bildwiederholraten** sind häufig die Folge von günstigen Kameras, geringem Speicherplatz, oder einer Infrastruktur, die nur eine geringe Bandbreite zur Verfügung stellen kann. Das stellt insbesondere bei Videoauswerteaufgaben, wie z.B. dem Tracking, eine Herausforderung dar, da dabei große Positionssprünge einer Person zwischen zwei aufeinanderfolgenden Bildern auftreten können.
- **Verschmutzungen** durch Staub, Regen oder auch Spinnennetze etc. stellen häufig weitere Herausforderungen im Videoüberwachungsbereich dar, die beispielsweise zu falschen Repräsentation von Personen im Bild führen können.
- **Schwache Kontraste** der Personen zum Hintergrund stellen im Rahmen dieser Arbeit hauptsächlich für das Tracking eine große Herausforderung dar. Ein *Verschwimmen* von Personen mit dem Hintergrund, aufgrund einer gleichen Kleiderfarbe wie der Hintergrund, führt oft zu einem schleichenden Übergang von der getrackten Person auf den Hintergrund oder zu einer Verwechslung mit einer anderen Person.
- **Verdeckungen**, wie beispielsweise verdeckte Körperteile, erschweren die Erstellung einer robusten Personenrepräsentation.
- **Ähnliche Erscheinungen** von Personen stellen für die erscheinungsbasierte Personenwiedererkennung eine große Schwierigkeit dar. Eine Differenzierung der Personen muss dann anhand sehr kleiner Farbunterschiede erfolgen.

Die meisten der genannten Herausforderungen stellen sich allen drei Bildauswerteaufgaben, allerdings unterschiedlich stark. Herausforderungen, die sich in der Regel aufgrund einer niedrigen Bildqualität stellen, sind in den Abbildungen 1.2, 1.3 und 1.4 dargestellt. In Abbildung 1.5 sind die anwendungsbezogenen Herausforderungen illustriert.

1.4 Beiträge

Gegenstand dieser Arbeit sind neuartige Methoden und Verbesserungen für die Detektion, Verfolgung und erscheinungsbasierte Wiedererkennung von

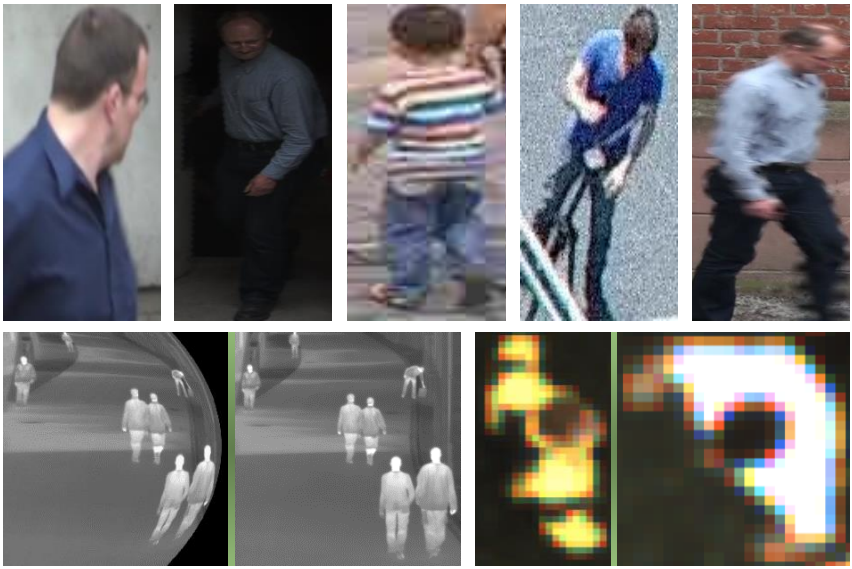


Abbildung 1.4: Beispiele für Herausforderungen bezüglich niedriger Bildqualität. In der ersten Reihe ist von links nach rechts jeweils ein Beispiel für Unschärfe, schlechte Beleuchtungsbedingung, Kompression, Bildrauschen und Zeilensprünge dargestellt. In der zweiten Zeile ist links ein Beispiel für die Linsenverzerrung illustriert und rechts sind zwei Bilder mit Farbsäumen, die an den Personenkonturen deutlich sichtbar sind, beispielhaft illustriert.

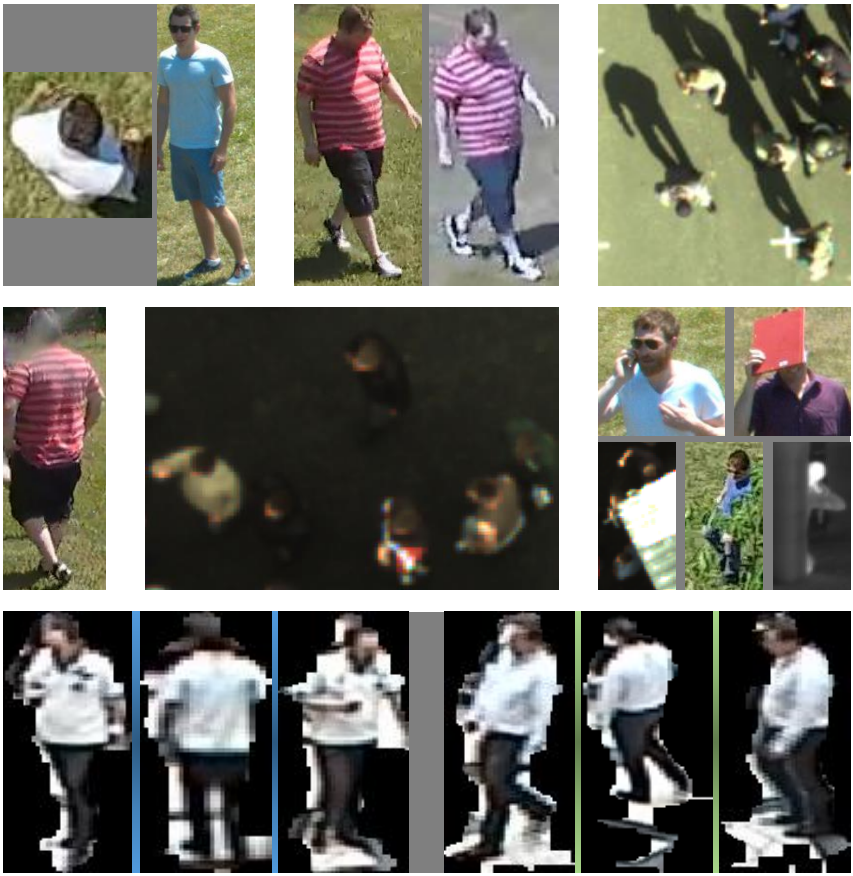


Abbildung 1.5: Illustration anwendungsbezogener Herausforderungen. Erste Reihe: Unterschiedliche Kameraausrichtungen, Farbunterschiede aufgrund von unterschiedlichen Kamerakonfigurationen und Schatten. Zweite Reihe: Verschmutzungen (im oberen Bereich des Bilds), schwacher Kontrast von Personen zum Hintergrund (die zwei schwarzgekleideten Personen in der Mitte) und Teilverdeckungen durch Gegenstände und Accessoires. In der untersten Reihe sind jeweils drei Bildausschnitte zweier verschiedener Personen zu sehen, wobei die zusammengehörenden Ausschnitte durch die farbigen Übergänge gekennzeichnet sind.

Personen, welche die oben genannten, im Bereich der Videoüberwachung typischen Herausforderungen berücksichtigen. Der Anwendungsschwerpunkt bildet dabei die Verarbeitung niedrig aufgelöster Bilder. Die Methoden basieren auf einem einheitlichen Bildauswerte-Rahmenwerk, das auf Kovarianzdeskriptoren [Tuz06] und einer riemannschen Mannigfaltigkeit von symmetrischen, positiv definiten Matrizen [Pen99, Pen06a, Pen06b, Pen06c] aufbaut. Der große Vorteil eines auf Kovarianzdeskriptoren basiertes Rahmenwerks ist seine Offenheit für beliebige Bildmerkmale, so dass Merkmale, die beispielsweise für die Detektion von Personen in Wärmebildern geeignet sind, einfach durch Merkmale für die farb- und erscheinungsbasierte Personenwiedererkennung ersetzt werden können. Sie können ohne Anpassungen der Verarbeitungsketten der in dieser Arbeit vorgestellten Bildauswerteaufgaben und des mathematischen Rahmenwerks in Kapitel 4 durch andere Merkmale ausgetauscht werden, weshalb im Rahmen dieser Arbeit die Auswahl der Merkmale eine untergeordnete Rolle spielt. Kovarianzdeskriptoren repräsentieren Merkmale durch eine Art Meta-Beschreibung, indem nicht die absoluten Werte der Merkmale betrachtet werden, sondern deren Varianzen und Korrelationen untereinander. Das Rahmenwerk hat sich im Rahmen dieser Arbeit für die Personenrepräsentation in niedrig aufgelösten Bildern sowie gegenüber den anderen Herausforderungen als geeignet erwiesen.

Die Methoden und Verbesserungen wurden in drei unterschiedlichen Anwendungsdomänen — Personendetektion, -tracking und -wiedererkennung — erarbeitet, um die Generalisierungsfähigkeit des zugrunde liegenden Bildauswerte-Rahmenwerks aufzuzeigen. Den Schwerpunkt bezüglich diesen Domänen bildet die erscheinungsbasierte Personenwiedererkennung in dieser Arbeit. Die konkreten Beiträge zu den jeweiligen Bildauswerteaufgaben dieser Dissertationsschrift sind im Folgenden zusammengefasst.

Personendetektion. Die Personendetektion verfolgt hier das Ziel, Personen in Einzelbildern anhand ihrer Körperteile (Körperteilmannigfaltigkeiten) zu detektieren. Der Kern des Verfahrens ist ein *Manifold Learning* (MaL)[2] Algorithmus, welcher Mannigfaltigkeiten einzelner Körperteile

[2]MaL ist eine Gruppe von Methoden zur nichtlinearen Dimensionsreduktion.

überwacht lernt, mit dem Ziel der Senkung der Intra-Variationen (Intra-Diskriminanz von Bildern einer Körperteilklasse) und der Erhöhung der Inter-Variationen zwischen unterschiedlichen Körperteilmannigfaltigkeiten (Inter-Diskriminanz).

Der Beitrag ist die

- Erweiterung eines kovarianzdeskriptorbasierten Personen- / Körperteildetektors mittels überwachter MaL Methode [Met10].

Personentracking. Das Ziel des Trackings im Rahmen der vorliegenden Arbeit ist in erster Linie die bildbasierte Verfolgung einzelner Personen in Menschenmengen. Es werden Anpassungen und Erweiterungen eines existierenden Kovarianz-Trackers vorgestellt.

Die Beiträge sind

- die Verbesserung eines Einzelpersonen-Trackingverfahrens mittels Anpassung eines kovarianzdeskriptorbasierten Personenmodells anhand statistischer Eigenschaften hinsichtlich der Mannigfaltigkeit der Kovarianzdeskriptoren [Met09] und
- die Erstellung eines Nadir-Datensatzes inkl. Evaluations-Metriken und Annotationen. Der Datensatz zeigt Personen und Personengruppen aus der Vogelperspektive mit unterschiedlichen Bewegungsdynamiken und ermöglicht die Erarbeitung und Evaluation von Trackingverfahren für niedrig aufgelöste Personen in (dichten) Personenmengen [Hüb08, Jäg08] (vgl. Abbildung 1.2).

Personenwiedererkennung. In diesem Teil der Arbeit soll eine Person, die manuell ausgewählt wurde, in Bild- oder Videoquellen unabhängig von Aufnahmeort und -zeit wiedergefunden werden. Es werden Tracklet[3]-Deskriptoren für die Repräsentation von Personen-Tracklets vorgestellt, die einen effizienten Vergleich ermöglichen. Darüber hinaus wird ein MaL basierter Optimierungsschritt zur Verbesserung der Wiedererkennungsergebnisse bzw. von Wiedererkennungsverfahren im Allgemeinen präsentiert.

[3] Von einem Trackingverfahren erzeugte zusammenhängende Bildsequenz.

Die Beiträge sind im Folgenden aufgelistet ([Met12b, Met12a]).

- Entwicklung einer neuartigen trackletbasierten Personenrepräsentation für die effiziente Überprüfung, ob mehrere Ganzkörpersequenzen von derselben Person stammen, um Personen kameraübergreifend wiedererkennen oder in Bilddatensätzen suchen zu können. Zugrunde liegen Kovarianzdeskriptoren aus denen mittels einer Spectral Clustering Strategie und einer Eigenwert-Analyse die Tracklet-Deskriptoren bestimmt werden.

- Erarbeitung einer MaL basierten Neusortierung von Teilergebnissen des Personenwiedererkennungsansatzes. Unter der Annahme, dass die zugrunde liegenden Kovarianzdeskriptoren auf einer Mannigfaltigkeit liegen, erfolgt eine einzelbildbasierte Neusortierung.

Die vorgestellten Methoden zur Detektion, Tracking und Wiedererkennung beruhen auf einem einheitlichen Bildauswerte-Rahmenwerk, das Bildausschnitte von Personen durch Kovarianzdeskriptoren repräsentiert. Neben diesen handentworfenen Merkmalen ist das Lernen von Merkmalen durch *tiefe künstliche faltende neuronale Netze* (TKFNN) ein sehr populärer anderer Ansatz. Die Arbeit schließt mit einem Anknüpfungspunkt zu den TKFNN und verdeutlicht

- anhand eines TKFNN, eine Möglichkeit, eine verbesserte kovarianzdeskriptorbasierte Repräsentation zu trainieren und

- anhand einer tiefen Fusion das Potential der Verknüpfung von Kovarianzdeskriptoren mit gelernten Merkmalen [Sch17].

Die Grundidee bei der verbesserten kovarianzdeskriptorbasierten Repräsentation ist es, ein TKFNN mit Bildausschnitten von Personen als Eingabe und vorberechneten Mittelwerten von Kovarianzdeskriptoren als Ausgabe zu trainieren, wobei die Mittelwerte aus Kovarianzdeskriptoren mehrerer Bildausschnitte berechnet werden. Bei der tiefen Fusion werden die handentworfenen Merkmale in den Trainingsprozess der zu lernenden Merkmale mit eingeschlossen, um die Stärken von handentworfenen und gelernten Merkmalen zu verknüpfen bzw. die Schwächen der einzelnen Merkmale zu kompensieren.

1.5 Gliederung

Die Dissertationsschrift gliedert sich wie folgt. Das folgende **Kapitel 2** gibt zunächst eine Übersicht über die für diese Arbeit relevante Literatur.

Anschließend werden in **Kapitel 3** Merkmale und Deskriptoren diskutiert, die für die Repräsentation von Personen in niedrig aufgelösten Bildern verwendet werden können. Das Hauptaugenmerk liegt dabei auf dem von Porikli et al. vorgestellten Kovarianzdeskriptor [Tuz06]. Dieser Deskriptor berechnet sich aus einfachen Merkmalen, wie z.B. Farbe oder Gradienten, und repräsentiert Einzelbildregionen anhand statistischer und örtlicher Eigenschaften der Merkmale bzw. Bildregion. Der Kovarianzdeskriptor wird für alle hier vorgestellten Verfahren verwendet, da er sich als besonders geeignet für die Repräsentation und den Vergleich von niedrig aufgelösten Personen erwiesen hat.

In **Kapitel 4** wird das auf den Arbeiten von Pennec et al. [Pen99, Pen06a, Pen06b, Pen06c] basierte mathematische Rahmenwerk zur Verarbeitung und Anwendung von Kovarianzdeskriptoren vorgestellt. Neben der riemannschen Mannigfaltigkeit positiv definiter Kovarianzmatrizen wird auch ein Ansatz für die nichtlineare Dimensionsreduktion beschrieben, der für diese Arbeit relevant ist.

Anschließend wird die Personendetektion anhand von Körperteilen in **Kapitel 5**, das Personentracking in Luftbildern mit dichten Menschenmengen in **Kapitel 6** sowie die erscheinungsbasierte Personenwiedererkennung in **Kapitel 7** vorgestellt. Die einzelnen Kapitel umfassen jeweils eine einleitende Darstellung der zugrunde liegende Aufgabe, die ausführliche Verfahrensbeschreibung sowie die Verfahrensevaluation und die Beschreibung der dafür verwendeten Datensätze.

Thematisch schließt die Arbeit in **Kapitel 8** mit einem Anknüpfungspunkt zu den TKFNN, indem Möglichkeiten aufgezeigt werden, eine verbesserte kovarianzdeskriptorbasierte Personenrepräsentation zu trainieren sowie handentworfene und gelernte Merkmalen so zu verknüpfen, dass eine Leistungssteigerung bei den Bildauswerteverfahren erreicht werden kann.

Den Schluss bildet das **Kapitel 9**, das die wesentlichen Ergebnisse dieser Arbeit zusammenfasst und einen Ausblick auf weiterführende Arbeiten gibt.

2

Stand der Forschung

Dieses Kapitel gibt einen Literaturüberblick über den Stand der Forschung, der für diese Arbeit relevant ist. Die Arbeiten werden entsprechend der Aufteilung dieser Dissertationsschrift gegliedert. Zunächst werden in Abschnitt 2.1 Ansätze zur Repräsentation von Personen in Videodaten mit niedriger Auflösung aufgeführt. Diese Repräsentationsmöglichkeiten fokussieren sich auf handentworfene, also nicht gelernte Merkmale. Anschließend an die Repräsentationsmöglichkeiten erfolgt in den Abschnitten 2.1.3 - 2.1.5 eine Übersicht über Arbeiten zu den in dieser Arbeit betrachteten Bildauswerteaufgaben Personendetektion, -tracking und -wiedererkennung.

In Abschnitt 2.1.6 werden relevante Arbeiten zu einer Repräsentationsart — Personenrepräsentation mittels Kovarianzdeskriptoren — aufgeführt, die im Rahmen dieser Arbeit eine wesentliche Rolle spielt. Kovarianzdeskriptoren eignen sich sowohl für Einzelbild- und bildsequenzbasierte Ansätze und haben über die letzten 10 Jahre breite Anwendung gefunden. Auch bezüglich den Herausforderungen in Abschnitt 1.3 erscheinen sie als geeignet. Eine genaue Betrachtung dieser Repräsentationsart hinsichtlich der Herausforderungen erfolgt in Kapitel 3.

Die Rolle gelernter Merkmale wird in Abschnitt 2.2 diskutiert. Es werden relevante Arbeiten bezüglich gelernter Merkmale für die Repräsentation bzw. Prozessierung von Personenbildern aufgeführt sowie in Abschnitt 2.3

Handentworfene Merkmale und Deskriptoren	Einzelbild	Bildsequenz	lokaler Ansatz	globaler Ansatz	Kontur	Erscheinung
SIFT [Low04]	×		×		×	
HOG [Dal05]	×			×	×	
LBP [Oja02]	×			×	×	×
ICF [Dol09]	×			×		×
Invariante Farbsignatur [Kvi13]	×			×		×
SDALF [Far10, Baz13]	×			×		×
Fisher-Vektoren [Per07]	×			×		×
Farbnamen[Yan14]	×			×		×
LOMO [Lia15]	×			×		×
Isomap [Ten00]		×		×	×	×
Laplacian Eigenmaps [Bel03]		×		×	×	×

Tabelle 2.1: Auswahl von handentworfenen Merkmalen und Deskriptoren, worauf relevante Personenrepräsentationen in der Regel aufbauen.

auch kombinierte Ansätze, die sowohl auf handentworfenen als auch auf gelernten Merkmalen basieren.

2.1 Personenrepräsentation mittels handentworfener Merkmale

Für die Repräsentation von Personen gibt es eine große Vielfalt an Merkmals- und Deskriptortypen [Rot08, Sat12]. Falls nicht explizit anders erwähnt, beziehen sich in dieser Arbeit Merkmale auf ein einzelnes Pixel

oder eine kleine Bildregion mit wenigen Pixeln und Deskriptoren, die sich im Allgemeinen aus den Merkmalen bestimmen, auf Bildausschnitte, die ganze Körperteile oder Personen repräsentieren. Die Tabelle 2.1 zeigt eine Auswahl relevanter handentworfener Merkmale und Deskriptoren, für sowohl einzelbild- und bildsequenzbasierte als auch kontur- und erscheinungslastige Repräsentationsansätze.

Bei handentworfenen Ansätzen ist die Personenrepräsentation ein wichtiger Schritt, wie z.B. in den Übersichtsartikeln [Rot08, Sat12, Xi13, Zhe16] deutlich wird. Im Folgenden werden die Merkmale und Deskriptoren aus der Tabelle 2.1 und weitere relevante Repräsentationsansätze kurz aufgeführt und bezüglich den Bildauswerteverfahren gruppiert, die im Rahmen dieser Arbeit betrachtet werden. Der Fokus liegt dabei, aufgrund der in dieser Arbeit betrachteten niedrigen Bildauflösungen, auf *einfachen* Merkmalen. Sowohl biometrische als auch attributbasierte und semantische Merkmale werden nicht betrachtet, da sie für diese Auflösungen im Allgemeinen ungeeignet sind. Die Gründe dafür werden in Kapitel 3 angesprochen, das die Eigenschaften bestimmter Merkmals- und Deskriptortypen näher thematisiert.

2.1.1 Einzelbildbasierte Personenrepräsentation

Eine Möglichkeit niedrig aufgelöste Personen in Einzelbildern zu repräsentieren ist die Berechnung von Modellen auf Basis lokaler Merkmale (Punktdeskriptoren). Ein bekannter Punktdeskriptor ist *Scale Invariant Feature Transform* (SIFT) [Low04] oder Erweiterungen davon, wie z.B. *Principal Component Analysis SIFT* (PCA-SIFT) [Yan04]. In der Dissertationsschrift [Jün11b] wurde er insbesondere erfolgreich für die Detektion und das Tracking von Personen in Ferninfrarotbildern eingesetzt. Eine andere Möglichkeit ist die Repräsentation von Personen durch bildregionenbasierte Deskriptoren (siehe z.B. [For07]), die in der Regel aus einfachen Merkmalen wie z.B. Kanten, Texturen und Farbe berechnet werden [Ngu16]. Einfache Merkmale sind *nah* an der Pixelinformation und beschreiben im Allgemeinen nur einzelne Pixelinformationen. Für die Personenrepräsentation werden aus den einfachen Merkmalen kontur- oder erscheinungsbasierte Deskriptoren bestimmt [Ngu16], die dann Bildausschnitte repräsentieren, die ganze Körperteile oder Personen zeigen. In

dem Übersichtsartikel zu Verfahren für die Personendetektion [Ngu16] ist eine gute Übersicht über verschiedene Ansätze gegeben, die nicht nur für Personendetektionsverfahren relevant sind.

Konturbasierte Repräsentationen konzentrieren sich auf die Kontur von Personen und werden primär für Detektions- (z.B. [Dal05]) und Trackingaufgaben verwendet (z.B. [Bil09]), wobei sie auch schon vorteilhaft für Wiedererkennungsverfahren eingesetzt wurden [Wan07]. Erscheinungsbasierte Deskriptoren hingegen beschreiben die Erscheinung einer Person in erster Linie anhand ihrer Kleiderfarben und werden oft durch einen Histogramm-Deskriptor repräsentiert, z.B. durch ein Farbhistogramm (z.B. [D'A11]) oder ein Histogramm mit visuellen Wörtern (z.B. [Wen11]). Da sich Personen farblich nicht immer von ihrer Umgebung unterscheiden, sind solche Deskriptoren nicht ausreichend diskriminativ für Detektionsaufgaben. Erscheinungsbasierte Deskriptoren sind eher für Wiedererkennungsaufgaben von Relevanz, können aber auch für Trackingaufgaben verwendet werden, insbesondere nach temporären Verdeckungen der zu trackenden Objekte. Deskriptoren können aber auch aus einer Kombination unterschiedlicher Merkmale entworfen werden [Ngu16], indem sie, wie z.B. in [Hah04], Farbe und strukturelle Informationen zusammen betrachten, was einem texturbasierten Deskriptor entspricht.

2.1.2 Bildsequenzbasierte Personenrepräsentation

Bei den bildsequenzbasierten Ansätzen werden die Personenrepräsentationen über mehrere, oft zeitlich zusammenhängende Bilder bzw. Deskriptoren berechnet. Die Bildsequenzen, bestehend aus ausgeschnittenen Personen, stammen dann üblicherweise von Personentrackingverfahren. Es können aber auch mehrere Bildausschnitte einer Person vorliegen, die von unterschiedlichen Quellen stammen, für die eine Repräsentation modelliert werden soll.

Eine der ersten populären bildsequenzbasierten Personenrepräsentationen wird in [Dal06] vorgestellt. Diese Autoren, die auch die *Histogramme orientierter Gradienten* (HOG) veröffentlicht haben, erweiterten die HOG um Histogramme, die auf dem optischen Fluss basieren. Ein anderer bekannter Ansatz ist in [Vio03] veröffentlicht, der auf den Unterschieden

von Pixelwerten zwischen zwei aufeinanderfolgenden Bildern aufbaut, die durch die Bewegung verursacht werden.

Weitere Ansätze, die sich in der Regel auf längere Bildsequenzen fokussieren, können in *mengen-*, *mittelwert-* und *modellbasierte* Repräsentationen eingeteilt werden.

- **Mengenbasierte** Repräsentationsarten sind eher für die Wiedererkennung interessant, wobei eine Person in der Regel durch mehrere Einzelbilder repräsentiert wird. Bei einem Person-zu-Person-Vergleich werden dann mehrere Deskriptoren einer Person mit mehreren Deskriptoren einer anderen Person verglichen. Dieser Ansatz wird auch als *Multi-Shot-Analyse* (MSA) bezeichnet. Erste solcher Ansätze für die Wiedererkennung sind in [Baz10, Far10] veröffentlicht.

- Bei den **mittelwertbasierten** Ansätzen werden mehrere Bilder durch wenige Mittelwerte oder sogar nur einen Mittelwert repräsentiert, die sich aus Teilsequenzen berechnen bzw. der sich aus der gesamten Bildsequenz berechnet. In [Bak12b] beispielsweise werden mehrere kleine aus Tracklets extrahierte Bildregionen gemittelt, die dann durch einen Mittelwert repräsentiert werden. Es können aber auch z.B. Kernels zeitlich gruppiert werden, um Bildsequenzen von Personen zu repräsentieren [Bau14].

- Die dritte Gruppe sind **modellbasierte** Ansätze. Eine Möglichkeit die Erscheinung modellbasiert zu repräsentieren ist mittels Hauptkomponentenanalyse [Jol86]. Dieses lineare Modell wird sehr oft — insbesondere im Bereich der Gesichtserkennung (siehe z.B. [Lev02, Elg08, Shr13, Kau15]) — verwendet, um Erscheinungen zu modellieren und Unterräume von Variationen der Erscheinungen zu entdecken [Elg08]. Andere Ansätze verwenden bilineare (z.B. [Ten00]) und multilineare (z.B. [Kar15]) Modelle, die jedoch auch eher für die Repräsentation von Gesichtern eingesetzt werden. Ein linearer Ansatz, der sich eher auf die Erscheinung des ganzen Körpers fokussiert, wird in [Kar15] vorgeschlagen. Bei diesem Ansatz wird versucht ein Anfragebild einer Person durch eine Linearkombination von Galeriebildern (Bilder in einer Datenbank) derselben Person zu repräsentieren, um eine Übereinstimmung festzustellen. Zeigen Bildsequenzen allerdings unterschiedliche Posen der Person, liegen die

Bilder in der Regel auf einer nichtlinearen Mannigfaltigkeit. Ein anderes anschauliches Beispiel ist die Betrachtung von Personenkonturen einer Person, die sich orthogonal zum Sichtstrahl der Kamera aufrecht fortbewegt. Fasst man die Konturen aus den einzelnen Bildern jeweils als einen Punkt auf, werden die Punkte aufgrund der Verformungen und Selbstverdeckungen auf einer nichtlinearen und gekrümmten Mannigfaltigkeit liegen, die durch eine Hauptkomponentenanalyse nicht gefunden werden kann [Elg08] (vgl. auch Abschnitt 4.2). Dafür sind entsprechende nichtlineare Modelle besser geeignet. Eine schöne Übersicht über Ansätze bis 2008 findet man in [Elg08] und neuere nichtlineare Ansätze sind z.B. [Con09, Tor10, Li15]. Dabei werden neben konventionellen Ansätzen der nichtlinearen Dimensionsreduktion zum Auffinden der zugrunde liegenden Mannigfaltigkeit, wie z.B. die verbreiteten MaL Algorithmen *Locally Linear Embedding* (LLE) [Row00], *Isometric Feature Mapping* (Isomap) [Ten00] oder *Laplacian Eigenmaps* (LE) [Bel03], auch neue Dimensionsreduktionsmethoden oder Erweiterungen der konventionellen Verfahren verwendet (z.B. [Li15]).

Im Rahmen dieser Arbeit wird ein nichtlinearer modellbasierter Ansatz für die Körperteildetektion (Kapitel 5) und mittelwertbasierte Ansätze für das Tracking (Kapitel 6) und die trackletbasierte Personenwiedererkennung (Kapitel 7) verwendet.

2.1.3 Detektion von Personen in Einzelbildern

Das Ziel der Personendetektion im Rahmen dieser Arbeit ist die Detektion von Personen in Wärmebildern anhand ihrer Körperteile. Das Verfahren entstand in erster Linie in Folge der Erarbeitung eines modellbasierten Ansatzes für die Personenrepräsentation auf Basis mehrerer Bilder, wobei die Detektion in Einzelbildern durchgeführt wird. Zudem soll der Detektionsansatz die Vorteile von Kovarianzdeskriptoren gegenüber einer direkten Verwendung einfacher Merkmale festigen (vgl. Abschnitt 3.4).

Generell können Personendetektionsverfahren in Ansätze auf Basis handentworfener und (überwacht) gelernter Merkmale gegliedert werden. Eine gute Übersicht über relevante Detektionsansätze ist in [Ngu16] zu finden.

Detektionsansätze, die auf handentworfenen Merkmalen basieren, werden hinsichtlich Detektionsleistung mittlerweile von Verfahren überholt, die auf überwacht gelernten Merkmalen basieren, wie z.B. die Ansätze in [Oli16, Liu16]. Für relevante Arbeiten auf Basis überwacht gelernter Merkmale wird an dieser Stelle auf den Abschnitt 2.2 verwiesen. Diese Verfahren müssen allerdings, im Vergleich zu den handentworfenen Ansätzen, (überwacht) gelernt werden, wofür sehr viele Bilder für das Training notwendig sind, um diese guten Ergebnisse erzielen zu können.

Bildmerkmale für die Personendetektion. Wie in [Ngu16] aufgeführt, waren HOG eine der ersten Personenrepräsentationen, die einige Jahre den Stand der Forschung im Anwendungsbereich der Personendetektion bildeten. Diese konturbasierten Deskriptoren wurden im Bereich der Personendetektion in zahlreichen Arbeiten mit verschiedenen Anpassungen und Erweiterungen verwendet (z.B. [Con13, Zha13a, Sat14]). Dass Repräsentationen, die auf Personenkonturen basieren, für Personendetektionsaufgaben prinzipiell gut geeignet sind, bestätigen auch psychologische Studien wie z.B. [dW04], die zeigt, dass solche Merkmale sehr gut diskriminierend sind, um Objekte in der menschlichen Wahrnehmung gut erkennen zu können (vgl. auch [Ngu16]).

Andere relevante Ansätze, wie z.B. [Yad08], berücksichtigen zudem Texturinformationen und verwenden dafür *Local Binary Pattern* (LBP) [Oja02]. In [Wan09, Zha11a] werden HOG und LBP und in [Hus10] werden HOG, LBP und *Local Ternary Pattern* (LTP) [Tan10] miteinander kombiniert. Dabei konnten mit der Dreierkombination teilweise bessere Ergebnisse als mit den einzelnen Merkmalen als auch mit Zweierkombinationen erzielt werden. Zudem wurde darin gezeigt, dass LBP und LTP komplementäre Eigenschaften besitzen und folglich entsprechend gut kombiniert werden können. Ein weiteres Ergebnis war, dass die LBP auf Farbbildern bessere Detektionsraten erzielten. Dass erscheinungsbasierte Merkmale — die eine größere Bedeutung bei Tracking- und Wiedererkennungsaufgaben haben — auch vorteilhaft für Detektionsaufgaben eingesetzt werden können, belegen auch andere Arbeiten. Zwei der ersten weit verbreiteten Ansätze, die die Erscheinung ergänzend mitberücksichtigen, sind in [Vio03, Dal06] veröffentlicht.

Weitere Ansätze beruhen auf sogenannten *Integral Channel Features* (ICF) [Dol09, Ben12], die Merkmale aus Integralbildern aufsummieren und dabei auch Farbe berücksichtigen. Farbansätze sind für Wärmebilder allerdings ungeeignet. Besser geeignete andere Ansätze, wie z.B. [Bou09], verwenden *Poselets*, die kleine Personenbereiche (Teile von Körperteilen ohne semantische Zuordnung) beschreiben. Poselets zeichnen sich durch eine hohe Invarianz gegenüber Orientierung und Betrachtungswinkel aus.

Aktuellere Ansätze setzen vermehrt auf *gelernte* Personenrepräsentationen. Die *Übergänge* von den handentworfenen zu den gelernten Repräsentationen in Abschnitt 2.2 stellen Verfahren dar, die auf *visuellen Codebüchern* basieren, wie z.B. die Ansätze in [Csu04, Lei06]. Die Verfahren beschreiben kleine Bildregionen durch sogenannte *visuelle Wörter*. Dabei werden zunächst häufig auftretende ähnliche kleine Bildregionen geclustert und dafür jeweils ein Clusterzentrum berechnet, das dann dem visuellen Wort entspricht (vgl. z.B. [Siv03]). Ein weiteres konkretes Beispiel für solch ein Ansatz, der eine Erweiterung der Codebuch-Ansätze darstellt und gute Ergebnisse auf den öffentlichen Personendetektionsdatensätzen *INRIA*[1] und *Caltech Pedestrian Detection Benchmark* [2] erreicht hat, ist in [Cos14] veröffentlicht.

2.1.4 Tracking von Personen in Menschenmengen

Beim Tracking ist das Ziel im Rahmen dieser Arbeit das robuste videobasierte Tracking einzelner Personen in Menschenmengen aus der Vogelperspektive. Grundsätzlich ist ein videobasiertes Trackingverfahren aus den Komponenten

- Objektdetektion,
- erscheinungsbasierte Modellierung/Repräsentation des Objekts für das Tracking,
- Assoziation des Objekts zwischen zwei Bildern und
- Bewegungs- und Positionsschätzung

[1] `http://pascal.inrialpes.fr/data/human/`
[2] `http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/`

aufgebaut [Xi13, Teu15]. In dieser Arbeit bilden die Schwerpunkte die Personenrepräsentation und die Zuordnung von Repräsentationen zwischen zeitlich aufeinanderfolgenden Videobildern. Die Assoziation der Personenrepräsentationen zwischen zwei Bildern wird — insbesondere bei Multi-Objekt-Trackingverfahren — oft als gewichtetes Zuordnungsproblem definiert, das durch die *Ungarische Methode* [Kuh55] gelöst wird, und in Kombination mit z.B. einem Kalman-Filter einen sehr robusten Multi-Objekt-Tracking-Ansatz bildet (siehe z.B. [Bew16, Sah16]). Weitere Assoziationsansätze sowie Bewegungs- und Positionsschätzungen sind in [Sme14, Gha15, Teu15] zu finden, die eine gute Gesamtübersicht über relevante Trackingverfahren geben. Für *Split-Merge*-Ansätze, die eher für Multi-Objekt-Tracking relevant sind und in dieser Arbeit ebenfalls eine untergeordnete Rolle spielen, wird auf die Dissertationsschriften [Teu15, Gri18] und die eigenen Arbeiten [Her11, Mül11] verwiesen. Ein guter Einstieg in dieses Thema stellt auch die Veröffentlichung [Hen11] dar.

Bildmerkmale für das Personentracking. Für das Tracking haben sich über Jahre hinweg Kombinationen aus kontur-, erscheinungs- und bewegungsbasierten Merkmalen als robuste Personenrepräsentationen erwiesen [Xi13]. In [Xi13] ist eine ausführliche Übersicht über verschiedene Repräsentationsmöglichkeiten für das Personentracking bis 2013 gegeben. An der grundsätzlichen Personenrepräsentation durch handentworfene Merkmale für Trackinganwendungen hat sich in den letzten Jahren nichts geändert. Viele Ansätze in [Xi13] haben sich bewährt und kommen nach [Mue17] in aktuellen Trackingverfahren immer noch zum Einsatz. Aktuelle erfolgreiche handentworfene Objektrepräsentationen für das Tracking einzelner Objekte sind Kombinationen aus HOG und (semantischen) Farbmerkmalen [Mue17, Cho17].

Relevante Trackingverfahren, wie beispielsweise der kernelbasierte Ansatz in [Har11], basieren teilweise sogar nur auf einfachen Merkmalen. Auch der weitverbreitete Ansatz in [Hen15], ein korrelationsfilterbasiertes Verfahren (*Kernelized Correlation Filters* (KCF)-Tracker), baut auf HOG auf. Korrelationsfilterbasierte Verfahren (*Korrelationstracker*) liefern seit dem *Minimum Output Sum of Squared Error* (MOSSE) Filter [Bol10] sehr gute Ergebnisse. Weitere aktuelle Trackingansätze, wie z.B. [Hon15] oder der Korrelationstracker in [Ma15], verwenden gelernte Merkmale, die

mittels tiefen künstlichen faltenden neuronalen Netzen gelernt werden. In [Ma15] werden dabei die Merkmale aus verschiedenen Schichten des Netzes extrahiert, damit sowohl örtliche Informationen für die Objektlokalisierung (vordere Schichten) als auch Invarianzen gegenüber Variationen der Erscheinungen (hintere Schichten) vorhanden sind. Es gibt auch Ansätze, die gelernte und handentworfene Merkmale erfolgreich kombinieren [Dan15].

2.1.5 Erscheinungsbasierte Personenwiedererkennung

Die erscheinungsbasierte Personenwiedererkennung ist ein jüngeres Forschungsfeld als die Personendetektion und -verfolgung [Mon11], weshalb der Fokus im Rahmen dieser Arbeit darauf liegt. In diesem Abschnitt werden die grundsätzlichen Ansätze hinsichtlich Prozessierung der Repräsentation ausführlicher betrachtet.

Die ersten erscheinungsbasierten Personenwiedererkennungsverfahren entstanden im Rahmen von Arbeiten im Bereich des Personentrackings in Multi-Kamera-Netzwerken (vgl. die Übersichtsartikel [Wan13, Zhe16]). Nach [Zhe16] wurde die Personenwiedererkennung das erste Mal explizit in der Arbeit [Zaj05] erwähnt. Das Ziel war die Wiedererkennung einer Person anhand von Farbe und trackletbasierter Bewegungsinformation, nachdem sie das Sichtfeld einer Kamera verlassen hat und wieder eingetreten ist. Ab 2006 wurden erste einzelbildbasierte Arbeiten ausschließlich zu dem Thema der Personenwiedererkennungsverfahren veröffentlicht: z.B. [Ghe06, Yoo06]. Diese beiden Verfahren bauen auf Bildausschnitten mit Personen auf, die mittels Segmentierungsverfahren anhand ihrer Bewegung aus Videos ausgeschnitten wurden. Die Verfahren verwenden Farb- und Strukturinformationen bzw. Abstände von Pixeln der Silhouetten zu den obersten Konturpixeln [Yoo06].

Erste bildsequenzbasierte Ansätze, die sich ausschließlich mit der Personenwiedererkennung beschäftigen und neben erscheinungs- und konturrelevanten Merkmalen auch Bewegungsinformationen berücksichtigen, wurden ab 2010 veröffentlicht [Zhe16]: Die Ansätze [Baz10, Far10] sind zwei erste Beispiele dafür, die u.a. durch die Berücksichtigung mehrerer Bildausschnitte einer Person eine Verbesserung gegenüber den einzelbildbasierten Verfahren erzielten.

Seit 2008 ist ein wachsendes Interesse an erscheinungsbasierten Personenwiedererkennungsverfahren zu beobachten [Gon14], das seit 2012 weiter rasant ansteigt [Zhe16]. Ein Grund für den rasanten Anstieg lässt sich vermutlich auf eine immer größer werdende Anzahl an (immer größeren) Kameranetzwerken sowie Bild- und Videodatenbanken zurückführen.

Relevante erscheinungsbasierte Wiedererkennungsverfahren auf Basis handentworfener Merkmale setzen den Schwerpunkt entweder eher auf den Entwurf der Merkmale oder auf das Lernen von Metriken (*Metric Learning* (MeL)) [Bak14]. Die Ansätze, die den Fokus auf die Merkmale legen (Abschnitt 2.1.5), also auf die Personenrepräsentation, verfolgen das Ziel, die Merkmale so zu entwerfen, dass sie möglichst umfassend die in Abschnitt 1.3 aufgeführten Herausforderungen lösen. Die zweite Gruppe von Verfahren hingegen *vernachlässigen* diesen Prozess und setzen auf MeL (Abschnitt 2.1.5), mit dem Ziel, die Variationen unterschiedlicher Erscheinungen einer Person (Intra-Variation) zu minimieren und die Variationen von Erscheinungen zwischen zwei unterschiedlichen Personen (Inter-Variation) zu maximieren.

Bildmerkmale für die Personenwiedererkennung. Erste erscheinungsbasierte Wiedererkennungsansätze basierten oft ausschließlich auf Farbinformationen (Farbhistogrammen) ohne eine örtliche Struktur zu berücksichtigen (siehe z.B. [Mad07]). Allerdings gibt es auch aktuellere Ansätze [Kvi13], die ausschließlich auf Farbe basieren. In [Kvi13] werden Farbmodelle verwendet, die sich insbesondere hinsichtlich Beleuchtungsänderungen als invariant erweisen, was eine grundsätzliche Herausforderung bei ausschließlicher Verwendung von Farbe darstellt.

Mittlerweile werden überwiegend Kombinationen aus Farb- und Textur- und/oder Konturmerkmalen verwendet. Erste Ansätze im Bereich der erscheinungsbasierten Personenwiedererkennung verknüpften beispielsweise Farbhistogramme mit Kanteninformationen, um Strukturen auf der Person zu beschreiben [Ghe06]. Ein anderes Beispiel für eine vielversprechende Kombination ist beispielsweise die Verknüpfung von Lab-Farbhistogrammen mit SIFT-Merkmalen [Zha13b]. In [Mig12] werden Farb- und Texturinformationen miteinander kombiniert: es werden Farbwerte aus drei Farbräumen mit LBP verknüpft. Ein Ansatz, der Personen durch

Farb-, Textur- und Konturmerkmale repräsentiert, ist in [Li14] veröffentlicht, wobei HSV-Farbhistogramme mit LBP und Gradientenhistogramme kombiniert werden.

In [Far10, Baz13] werden symmetrische Eigenschaften von Personen bei der Repräsentation mitberücksichtigt. Farenzena & Bazzani verfolgen den Ansatz — der auch unter dem Akronym *Symmetry-Driven Accumulation of Local Features* (SDALF) bekannt ist — Merkmale, die nah an der vertikalen Symmetrieachse liegen, stärker zu gewichten als Merkmale, die weiter weg sind. Damit soll sichergestellt werden, dass nur Merkmale der Person in die Repräsentation mit einfließen. Dieses Verfahren hat sich als geeignet für niedrig aufgelöste Personen und robust gegenüber Teilverdeckungen, Posenvariationen sowie Blick- und Beleuchtungsänderungen erwiesen [Far10, Baz13]. Außerdem werden bei diesem Verfahren die Farbinformation über die Zeit akkumuliert, um eine bessere Farbrepräsentation zu erhalten, weshalb dieser Ansatz auch zu den bildsequenzbasierten Ansätzen zugeordnet werden kann (vgl. Abschnitt 2.1.2).

In [Ma12b] wird eine neuere Repräsentation vorgeschlagen, die auf *Fisher-Vektoren* [Per07] basiert, mit der noch bessere Wiedererkennungsergebnisse erzielt werden konnten. Bei diesem Repräsentationsansatz werden in kleinen Bereichen innerhalb des zu repräsentierenden Bildausschnitts aus einfachen Merkmalen lokale Deskriptoren berechnet und zu Fisher-Vektoren transformiert, so dass sich die globale Repräsentation durch eine Menge solcher Vektoren bestimmt. Ein weiterer aktueller handentworfener Deskriptor, mit dem gute Ergebnisse erzielt wurden und der in einigen erscheinungsbasierten Wiedererkennungsverfahren eingesetzt wird [Zhe16], ist der *Local Maximal Occurrence* (LOMO)-Deskriptor [Lia15]. Dabei werden Auftrittwahrscheinlichkeiten von Merkmalen innerhalb mehrerer kleiner Bildausschnitte eines horizontalen Bildstreifens analysiert. Damit sollen ähnlich dem Ansatz in [Liu12] mittels seperater Betrachung horizontaler Bildausschnitte eine größere Invarianz gegenüber unterschiedlicher Betrachtungswinkel erreicht werden. Um eine Beleuchtungsinvarianz zu erlangen, werden außerdem die Eingabebilder (Personenbildausschnitte) mittels dem *Retinex*-Algorithmus vorverarbeitet und es wird ein zusätzlicher skalierungsinvarianter Texturoperator verwendet. Kombiniert mit MeL konnten damit Stand-der-Forschung-Verfahren auf vier öffentlichen Personenwiedererkennungsdatensätzen geschlagen werden (MeL wird in Abschnitt 2.1.5 behandelt).

In [Yan14] wird ein Deskriptor vorgeschlagen, der die Personen auf semantische Weise repräsentiert. Die Idee dahinter ist die Verwendung von *Farbnamen*, die sich aus Farbmerkmalen mehrerer Farbräume bestimmen. Durch die Verwendung von Farbnamen anstatt den Merkmalen direkt, konnten sie Stand-der-Forschung-Verfahren auf zwei relevanten öffentlichen Datensätzen übertreffen. Weitere Ansätze, die in diese Richtung gehen, sind attributbasierte Ansätze. Eine gute Übersicht solcher Verfahren ist in der Übersichtspublikation [Zhe16] gegeben.

Erhöhung der Diskriminanz. In der Personenwiedererkennung ist es wichtig, dass Merkmale und Deskriptoren diskriminativ sind. Ein Ansatz die Diskriminanz zu erhöhen ist *Metric Learning* (MeL). Die Idee hinter MeL ist es anhand von Personenlabels eine Pseudometrik zu lernen, so dass Personenrepräsentationen einer Person möglichst — auch bei großen Variationen in der Erscheinung — *nah zusammenrücken* und Personenrepräsentationen unterschiedlicher Personen sich möglichst *weit voneinander entfernen*. Sehr ausführliche Übersichten über MeL-Methoden sind in [Yan06, Zhe16] zu finden.

Viele MeL-Methoden basieren auf der Mahalanobis-Distanz [Zhe16], die wohl bekannteste MeL-Methode ist KISSME [Kos12]. Angenommen die Personenrepräsentation basiert auf Merkmalsvektoren, dann verfolgen diese Verfahren das Ziel, den Merkmalsraum so linear zu transformieren, dass relevante Merkmalsdimensionen höher als irrelevante Dimensionen gewichtet werden. Solche MeL-Methoden können für Merkmale direkt oder auf dimensionsreduzierte Merkmale angewandt werden, mit dem Ziel für jede Person einen Unterraum zu lernen, der zu den anderen möglichst diskriminativ ist [Lia15].

Andere Ansätze, die nicht auf der Mahalanobis-Distanz basieren, wie beispielsweise [Zha16] oder [Gra08], verwenden *Support-Vektor-Maschinen* (SVM) [Sch02] oder *Boosting*-Methoden [Vio01, Tie04]. Mittels diesen konventionellen Klassifikationsverfahren werden dann die einzelnen Merkmalsdimensionen bzw. Merkmale gewichtet.

Weitere Ansätze erhöhen die Diskriminanz während der Dimensionsreduktion, indem sie diskriminative Merkmalsunterräume lernen [Zhe16].

In [Ped13] beispielsweise wird ein zweistufiger Dimensionsreduktionsansatz gewählt, der zunächst mittels Hauptkomponentenanalyse und einer anschließenden linearen Fisher'schen Diskriminanzanalyse Unterräume bestimmt, in denen die Merkmale unterschiedlicher Personen besser trennbar sind. Ein anderer, ähnlicher Ansatz wurde bereits in Zusammenhang mit MeL erwähnt. Liao et al. bestimmen zunächst ähnlich wie in [Ped13] diskriminative Unterräume, worin sie zusätzlich MeL anwenden [Lia15].

2.1.6 Detektion, Tracking und Wiedererkennung mit Kovarianzdeskriptoren

In diesem Abschnitt werden relevante Arbeiten zu Kovarianzdeskriptoren betrachtet, die in [Tuz06] für die Objektdetektion und Texturklassifikation vorgeschlagen wurden, wobei sie auch für andere Bildauswerteverfahren geeignet sind (vgl. Kapitel 3). Sie können in Abhängigkeit der Zieldefinition der Anwendung bzw. der Bilddaten z.B. kontur- oder erscheinungsbasiert definiert oder aus einer beliebigen Kombination beider Merkmalstypen aufgebaut werden. Kovarianzdeskriptoren für Detektionsverfahren können beispielsweise aus LBP-Merkmalen berechnet werden und für Trackinganwendungen kann z.B. eine Kombination aus Farb- und Gradienteninformationen zugrunde gelegt werden. Es können auch Bewegungsinformationen mitberücksichtigt werden, womit Kovarianzdeskriptoren auch für bildsequenzbasierte Personenrepräsentationen geeignet sind, oder aber auch beliebig andere und folglich auch zukünftige Merkmale verwendet werden, die die Leistung der Merkmale des aktuellen Forschungsstands übertreffen. Die Kovarianzdeskriptoren sollten dann erwartungsgemäß allerdings besser oder zumindest ähnlich gut wie die Merkmale abschneiden. Der letzte Punkt wird im folgenden Kapitel nochmal aufgegriffen und näher erläutert.

Aufgrund der Repräsentation durch Kovarianzdeskriptoren werden die Merkmale nicht mehr direkt für die Repräsentation verwendet, sondern mittels Kovarianzmatrizen repräsentiert. Die Matrizen beschreiben damit die Varianzen der Merkmale und die Korrelationen zwischen den Merkmalen, was — wie in Kapitel 3 verdeutlicht wird — verschiedene Vorteile bietet, die wiederum ausschlaggebend für die Fokussierung auf Kovarianzdeskriptoren im Rahmen dieser Arbeit waren. Eine ausführliche

Vorstellung dieser Deskriptoren, auch hinsichtlich den Herausforderungen in Abschnitt 1.3, wird in Kapitel 3 gegeben.

Es gibt zahlreiche Anwendungsdomänen in denen Kovarianzdeskriptoren erfolgreich eingesetzt werden. Neben den Bildauswerteaufgaben, die für diese Arbeit relevant sind, werden sie zudem für z.B. den Vergleich und die Suche von 3D-Formen [Tab14], die Erkennung von Aktivitäten und Gesten [San13] sowie die Gesichtswiedererkennung [Wan12] verwendet.

Im Folgenden werden relevante Arbeiten betrachtet, die Kovarianzdeskriptoren für die Personendetektion, das Personentracking oder die erscheinungsbasierte Personenwiedererkennung einsetzen. Eine Übersicht über diese Arbeiten ist in den Tabellen 2.2 und 2.3 gegeben.

Personendetektion mit Kovarianzdeskriptoren

In der Publikation [Tuz06], in der die Kovarianzdeskriptoren vorgestellt wurden, wurden sie für die Detektion von Objekten und Texturklassifikation vorgeschlagen. Darin wurde qualitativ gezeigt, dass sie bessere Detektionsergebnisse als andere Merkmale bzw. Deskriptoren erreichen können, die dem damaligen Stand der Forschung entsprachen. Quantitative Ergebnisse, die das bestätigen, haben Tuzel et al. in [Tuz07] veröffentlicht. Darin verwenden Tuzel et al. einen Boosting-Ansatz zur Klassifikation der Deskriptoren, also zur Bestimmung, ob der Deskriptor eine Person zeigt oder nicht, und erzielten damit bessere Ergebnisse als [Dal05, Zhu06]. In [Lia11] erzielten Liao & Huang unter selben Trainingsbedingungen mit Kovarianzdeskriptoren auf drei öffentlichen Datensätzen auch bessere Ergebnisse als mit HOG. Dabei muss allerdings berücksichtigt werden, dass sie für die Klassifikation von Kovarianzdeskriptoren eine Kombination aus dem *naiven Bayes-Klassifikator* und einer Kaskade von SVM verwendet und bei dem HOG-Ansatz *adaptives Boosting* zur Klassifikation angewandt haben.

In [Yao08] wurden Erweiterungen des Detektionsansatzes von Tuzel et al. vorgestellt. Bei diesem Ansatz werden u.a. Bewegungsinformationen als Merkmale berücksichtigt und aus Laufzeitgründen mittels Merkmalsselektion diskriminative Untermengen von Merkmalen bestimmt, woraus die Kovarianzdeskriptoren berechnet werden. Für jede Untermenge wird ein

Veröffentlichung	Personendetektion	Personentracking
Tuzel et al. [Tuz06]	×	
Tuzel et al. [Tuz07]	×	
Yao & Odobez [Yao08]	×	
Tosato et al. [Tos10]	×	
Liao & Huang [Lia11]	×	
Yao & Odobez [Yao11]	×	
Porikli et al. [Por06b]		×
Li et al. [Li08]		×
Palaio & Batista [Pal08]		×
Tyagi et al. [Tya08]		×
Wu et al. [Wu08]		×
Wu et al. [Wu09]		×
Hong et al. [Hon10]		×
Wu et al. [Wu12a]		×
Wu et al. [Wu12b]		×
Hu et al. [Hu12]		×
Wu et al. [Wu15]		×

Tabelle 2.2: Übersicht über relevante kovarianzdeskriptorbasierte Arbeiten zu Personendetektion und -tracking.

eigener Klassifikator bzw. eine Kaskade mehrerer schwacher Klassifikatoren (*Boosting-Klassifikator*) trainiert. In [Yao11] haben diese Autoren weitere Verbesserungen in Bezug auf ihre Arbeit in [Yao08] vorgestellt, wobei diese weniger die Kovarianzdeskriptoren selbst, sondern vielmehr Vorverarbeitungsschritte betrifft, wie z.B. Rektifizierung von Bildregionen oder die Fokussierung auf Vordergrundmerkmalen (Merkmale auf den Personen).

Bei den genannten Detektionsansätzen lag der Schwerpunkt auf der Repräsentation von Bildregionen, die eine Person zeigen. Dabei wird die Bildregion durch einen Kovarianzdeskriptor repräsentiert. In [Tos10] hingegen werden solche Bildregionen hierarchisch unterteilt. Es werden insgesamt 11 kleinere Bildregionen darin bestimmt, die sowohl den ganzen Körper als auch Körperregionen wie z.B. Kopf-Schulter-Partien und einzelne Körperteile wie z.B. den *rechten Arm* repräsentieren, wofür jeweils Kovarianzdeskriptoren berechnet werden. Die einzelnen Deskriptoren werden dann mittels Boosting-Klassifikator klassifiziert, wobei für jedes Körperteil bzw. jede Körperregion und auch für den *ganzen Körper* ein separater Boosting-Klassifikator gelernt werden muss.

Kovarianzdeskriptorbasiertes Tracking einzelner Personen

In [Por06b] verwenden Porikli et al., welche die Kovarianzdeskriptoren für die Repräsentation von Bildregionen in [Tuz06] vorgeschlagen haben, die Deskriptoren zum Tracking einzelner Objekte. Dabei werden Kovarianzdeskriptoren auf Basis von Gradienten- und Farbmerkmalen verwendet, um das Assoziationsproblem zu lösen. Das Verfahren eignet sich sowohl für starre als auch nicht starre Objekte und kann auch auf Bildsequenzen eingesetzt werden, die mit einer bewegten Kamera akquiriert wurden.

In [Pal08, Wu08, Wu12a] wurde der Ansatz von Porikli et al. aufgegriffen und hinsichtlich Bewegungs- und Positionsschätzung verbessert. Darin werden Partikelfilter auf die riemannsche Mannigfaltigkeit der Kovarianzdeskriptoren angewandt, um die Robustheit bzgl. Störungen im Hintergrund zu verbessern. In [Tya08] werden weitere Ergänzungen bzw. Anpassungen vorgestellt, sowohl zur Verbesserung der Robustheit als auch der Laufzeit des Porikli'schen Kovarianz-Trackers. Die Robustheit konnten sie u.a. durch den *Mean Shift* Algorithmus in [Com02] verbessern und die Laufzeit

konnten Tyagi et al. steigern, indem nicht alle Abstände in einer Pixelumgebung berechnet werden, sondern mittels Gradientenabstiegsverfahren sich dem *besten Match* angenähert wird. In den genannten Ansätzen basieren die Kovarianzdeskriptoren auf Merkmalen, die u.a. nicht rotationsinvariant sind. Um eine Invarianz gegenüber z.B. Rotation zu erlangen, können rotationsinvariante Merkmale definiert werden. Eine andere Möglichkeit wird in [Wu15] vorgestellt. Darin wird die Skalierung und Orientierung mittels affinen Kernel-Funktionen mitgeschätzt und im Assoziationsschritt berücksichtigt.

Weitere Ansätze mit Kovarianzdeskriptoren sind in [Li08, Wu09, Hon10, Wu12b, Hu12] veröffentlicht, die allerdings — aus praktischer Sicht — eine einfachere Abstandsfunktion als die riemannsche Metrik, die die Autoren in [Por06b] verwenden, einsetzen. Die genannten Ansätze verwenden die sogenannte *log-euklidische Metrik* [Ars06], die eine Verarbeitung der Kovarianzdeskriptoren im euklidischen Raum ermöglicht. Dabei werden die logarithmierten Kovarianzdeskriptoren betrachtet, die auch als *log-euklidische Kovarianzmatrizen* bekannt sind. Die Ansätze verfolgen alle den ähnlichen Ansatz, aus den log-euklidische Kovarianzmatrizen über das Tracking hinweg eine diskriminative und robuste Repräsentation für die Erscheinung der getrackten Person zu lernen. In [Li08, Wu09, Wu12b, Hu12] werden dazu log-euklidische Unterräume für die Repräsentation gelernt und in [Hon10] wird eine *Ein-Klassen-SVM* für die Repräsentation angewandt. In Kapitel 4, in dem ein auf den Arbeiten in [Pen99, Pen06a, Pen06b, Pen06c] basiertes riemannsches Rahmenwerk für die Verarbeitung von Kovarianzdeskriptoren beschrieben ist, wird darüber hinaus auch diese Metrik kurz erläutert.

Ähnliche kovarianzdeskriptorbasierte Repräsentationen aus kontur- und erscheinungsbasierten Merkmalen wurden auch bei erscheinungsbasierten Wiedererkennungsverfahren aufgegriffen, die im folgenden Abschnitt behandelt werden.

Erscheinungsbasierte Wiedererkennung mit Kovarianzdeskriptoren

Nach [Bak14] wurden Kovarianzdeskriptoren in [Bak10] das erste Mal für die erscheinungsbasierte Personenwiedererkennung zwischen unterschiedlichen Kameras, deren Sichtbereiche sich nicht überlappen, verwendet. Die

Veröffentlichung	**Erscheinungsbasierte Personen-wiedererkennung**	**Metric Learning**	**Dimensionsreduktion**
Bąk & Brémond et al. [Bak10]	×		
Hirzer et al. [Hir11]	×		
Zhang & Li [Zha11b]	×		
Ayedi et al. [Aye12]	×		
Bąk et al. [Bak12a]	×		
Bąk et al. [Bak12b]	×		
Li & Wang [Li12a]	×		
Ma et al. [Ma12a]	×		
Bąk et al. [Bak14]	×	×	
Eiselein et al. [Eis14]	×		
Serra et al. [Ser14]	×		
Matsukawa et al. [Mat16]	×		
Sivalingam et al. [Siv09]		×	
Vemulapalli & Jacobs [Vem15]		×	
Matsuzawa et al. [Mat17]		×	
Horev et al. [Hor17]			×

Tabelle 2.3: Übersicht über relevante kovarianzdeskriptorbasierte Arbeiten zu Personenwiedererkennung (erscheinungsbasiert) sowie relevante Veröffentlichungen zu MeL und Dimensionsreduktion für Kovarianzdeskriptoren.

Wiedererkennung wird anhand von bis zu sechs Körperteilen durchgeführt, die mittels HOG detektiert wurden. Für die detektierten Körperteile werden Kovarianzdeskriptoren aus Farb- und Gradientenmerkmalen sowie Pixelkoordinaten (für die örtliche Struktur) berechnet und mittels einem pyramidalen Abgleich miteinander verglichen. Die zugrunde liegende Metrik ist die in Abschnitt 4 aufgeführte riemannsche Metrik (Gleichung (4.1.1)).

In [Hir11] wird ein zweistufiger Ansatz verfolgt: Zunächst erfolgt ähnlich wie in [Bak10] ein Abgleich des Anfragebilds mit Bildern aus einem zu durchsuchenden Datensatz, wofür auch Kovarianzdeskriptoren auf Basis von Farb- und Gradientenmerkmalen sowie Pixelkoordinaten verwendet werden. Statt auf Körperteilen werden die Deskriptoren allerdings auf sieben Bildregionen berechnet, die sich durch Aufteilung der gesamten Bildregion (Person) in gleichgroße horizontale Streifen ergeben. Das Ergebnis nach der ersten Stufe ist eine Liste von Galeriebildern, die nach der Ähnlichkeit zum Anfragebild sortiert sind. Anschließend erfolgt eine optionale Neusortierung der Liste mittels MSA. Dafür wird mit Haarähnlichen Merkmalen [Vio01] und aus Kovarianzdeskriptoren berechneten *Sigma-Punkten* [Jul96, Klu10] ein diskriminatives Modell gelernt (Boosting-Klassifikator).

In [Bak12b] werden ebenfalls Kovarianzdeskriptoren für die Personenrepräsentation verwendet. Die Wiedererkennungsaufgabe wird darin als Klassifikationsaufgabe definiert, die durch Boosting gelöst wird. Bei der Klassifikation von Kovarianzdeskriptoren muss berücksichtigt werden, dass sie in keinem euklidischen Raum, sondern auf einer riemannschen Mannigfaltigkeit liegen (vgl. Kapitel 4). In [Bak10] wird die Klassifikation deshalb auf den Tangentialräumen, und in [Hir11] werden aus den Kovarianzdeskriptoren Sigma-Punkte berechnet, um eine euklidische Beschreibung von Kovarianzdeskriptoren zu erhalten. Nach [Bak14] liefern solche diskriminativen Ansätze zwar gute Ergebnisse, sind allerdings auch sehr rechenintensiv und nur eingeschränkt skalierbar. Mit jeder neuen Person die wiedererkannt werden soll, muss der Klassifikator neu und auf einer größeren Datenmenge trainiert werden, damit eine akzeptable Diskriminanz zwischen den Personen erreicht werden kann. Hirzer et al. haben dafür in [Hir11] und Bąk et al. in [Bak11, Bak12a] Ansätze zur Merkmalsselektion vorgestellt, die aufgrund einer geringeren Merkmalsmenge ein effizienteres Training ermöglichen und einer Überanpassung des Klassifikators beim

Training entgegenwirken können. Die Skalierbarkeit bleibt aber dennoch eingeschränkt.

Ein anderer Ansatz für die Personenrepräsentation für Wiedererkennungsaufgaben auf Basis von Kovarianzdeskriptoren wird in [Ma12a] vorgeschlagen. Dabei ergibt sich die Personenrepräsentation nicht direkt aus Kovarianzdeskriptoren, die aus Merkmalen bestimmt wurden, sondern aus den Abständen zwischen mehreren verschiedenen Kovarianzdeskriptoren, die innerhalb eines Bildausschnitts einer Person berechnet wurden. Die einzelnen zugrunde liegenden Kovarianzdeskriptoren basieren im Wesentlichen auf Farbe und *Gabor*-Merkmale (Merkmale aus 2D-Gabor-gefilterten Bildern), die u.a. die Invarianz gegenüber Beleuchtungsvariationen weiter verbessern können. Grundsätzlich bieten die Kovarianzdeskriptoren schon eine Invarianz bzgl. Beleuchtungsvariationen (siehe Kapitel 3). Neben guten Ergebnissen in der erscheinungsbasierten Personenwiedererkennung, konnten auch gute Resultate im Bereich der Gesichtswiedererkennung und -verifizierung erzielt werden.

In [Zha11b] konnten davor schon gute Ergebnisse mit Kovarianzdeskriptoren auf Basis von Farbe und Gabor-Merkmalen erzielt werden. Dabei werden die Kovarianzdeskriptoren direkt aus den Merkmalen berechnet, wobei neben Farbe und Gabor-Merkmale auch LBP verwendet werden. Zhang et al. haben LBP hinzugefügt, mit dem Ziel, die Beleuchtungsvariationen noch besser zu kompensieren. Auch in [Eis14] werden mit dem Ziel bessere Wiedererkennungsraten zu erzielen mehrere komplementäre Merkmale berücksichtigt und mittels Kovarianzdeskriptoren miteinander kombiniert. Es existieren einige Ansätze — speziell für Kovarianzdeskriptoren — die mittels MeL versuchen, die Diskriminanz der Deskriptoren und folglich die Wiedererkennungsleistung zu verbessern (siehe z.B. [Siv09, Vem15, Mat17]). Andere Ansätze zur Erhöhung der Diskriminanz sind nichtlineare Dimensionsreduktionsverfahren, wie z.B.[Hor17], der die Mannigfaltigkeit der positiv definiten Kovarianzdeskriptoren bei der Dimensionsreduktion mit berücksichtigt.

Weitere Ansätze, die auch über den Bereich der Personenwiedererkennung hinausreichen, berechnen für die Personenrepräsentation Modelle, die sich aus mehreren Kovarianzdeskriptoren bestimmen (z.B. [Bak10, Aye12, Li12a, Ser14, Mat16]). Solche Ansätze sind insbesondere für die Repräsentation höch aufgelöster Personen bzw. Bildausschnitte im Allgemeinen

interessant. Bei der Repräsentation einer hoch aufgelösten Person durch einen einzigen Kovarianzdeskriptor, würden lokale — ggf. diskriminative — Strukturen der Person verschwinden.

2.2 Personenrepräsentation mittels gelernter Merkmale

Die Verwendung überwacht gelernter Merkmale für die Personenrepräsentation ist seit Anfang der 2010er gestiegen, wie beispielsweise in [Sri16, Xia16] zu sehen ist. Das ist u.a. die Folge heutiger Möglichkeiten, sehr tiefe künstliche neuronale Netze in akzeptabler Zeit zu trainieren und der daraus resultierenden guten Ergebnisse in der Mustererkennung. Für eine umfassende Einführung in (tiefe) künstliche neuronale Netze wird auf [Goo16] verwiesen.

Tiefe künstliche neuronale Netz, wie z.B. das Konvolutionsnetz in [Fuk80], gibt es seit den 80er Jahren und erleben seit Anfang der 2010er eine Renaissance [Raw17]. Im Bereich der Bildauswertung bzw. Mustererkennung haben tiefe künstliche faltende neuronale Netze (TKFNN) in dem *ImageNet*-Wettbewerb [Den09] ihren Durchbruch erreicht, bei der Einzelbilder anhand der abgebildeten Objekte in eine von 1000 Objektklassen zugeordnet werden mussten. Ein auf diesem Wettbewerb eingereichter TKFNN-Ansatz ist AlexNet [Kri12], der ein deutlich besseres Klassifikationsergebnis als konkurrierende und andere vorher veröffentlichte Ansätze erreicht hat.

In der Tabelle 2.4 ist eine Auswahl der wichtigsten TKFNN-Ansätze gegeben.

Im Bereich der Detektion von Körperteilen erzielen Detektionsverfahren wie z.B. [Tia15, Vu15] sehr gute Ergebnisse, die auf dem *R-TKFNN*-Ansatz (Regionen mit TKFNN-Merkmalen) in [Gir14] basieren. Generell besteht dieser Ansatz aus zwei Schritten: Bestimmung von Bildregionen (*regions*), die möglicherweise ein Objekt zeigen, und Klassifikation der Regionen anhand daraus extrahierter TKFNN-Merkmalen mittels konventioneller Klassifikatoren wie z.B. SVM. Die Bildregionen werden mit TKFNN oder anderen speziellen *Region Proposal* Ansätzen bestimmt (vgl. [Gir14]).

Veröffentlichung	**Detektion von Körperteilen**	**Tracking einzelner Personen**	**Erscheinungsbasierte Wiedererkennung (Verifikationsansatz)**	**Erscheinungsbasierte Wiedererkennung (Identifikationsansatz)**
Girshick et al. [Gir14]	×			
Tian et al. [Tia15]	×			
Vu et al. [Vu15]	×			
Nam & Han [Nam16]		×		
Danelljan et al. [Dan16]		×		
Gordon et al. [Gor17]		×		
Yi et al. [Yi14]			×	
Li et al. [Li14]			×	
Varior et al. [Var16]			×	
Wu et al. [Wu16]			×	
Chen et al. [Che17]			×	
McLaughlin et al. [McL16]				×
Krizhevsky et al. [Kri12]				×
Simonyan & Zisserman [Sim14]				×
He et al. [He16]				×
Xiao et al. [Xia16]				×

Tabelle 2.4: Übersicht über verschiedene TKFNN-Ansätze, die bezüglich Bildauswerteaufgabe bzw. Verfahrensansatz gruppiert sind.

Im Bereich des Trackings einzelner Personen erzielen sogenannte *hybride* Ansätze sehr gute Ergebnisse (vgl. [Kri15, Kri16]). Bei diesen Verfahren wird in der Regel zunächst offline, ohne Bilder von den zu trackenden Objekten zu verwenden, eine generische Personenrepräsentation mittels TKFNN gelernt und online, also zur Laufzeit des Trackings an die Erscheinung des zu trackenden Objekts angepasst (siehe z.B. [Nam16, Dan16]). Der *Multi-Domain-Network*-Tracker aus [Nam16] gewann den Tracking-Wettbewerb *Visual Object Tracking VOT2015 Challenge* (VOT2015) [Kri15] und der *C-COT*-Tracker (Continuous Convolution Operator Tracker) aus [Dan16] den VOT-Wettbewerb ein Jahr später (*Visual Object Tracking VOT2016 Challenge* (VOT2016)) [Kri16]. Ein anderer vielversprechender Ansatz ist [Gor17], wobei ein rekurrentes neuronales Netz zum Tracking verwendet wird. Nach den Autoren ist es der erste Algorithmus, der solche Netze für das Tracking einzelner generischer Objekte in vielfältigen natürlichen Sequenzen und Situation verwendet [Gor17]. Das Verfahren liefert ähnlich gute Ergebnisse wie die jeweils 10 besten Tracker aus den Wettbewerben VOT2015 und VOT2016, wobei dieser Ansatz schneller als die meisten anderen ist und damit aus praktischer Sicht sehr interessant ist.

Erste vielversprechende TKFNN-Ansätze für die erscheinungsbasierte Wiedererkennung von Personen, wie z.B. [Yi14, Li14], verwenden *siamesische* neuronale Netze [Bro93], um zu bestimmen, ob zwei Eingabebilder dieselbe Person zeigen [Zhe16]. Auch neuere Verfahren, wie beispielsweise [Var16, Wu16], die bessere Ergebnisse erzielen, verfolgen ähnliche Ansätze. Ein ähnlicher Ansatz, der mehr als zwei Anfragebilder betrachtet, ist [Che17]. Mittlerweile übertreffen allerdings Ansätze, die sich auf die Personenrepräsentation mittels TKFNN-Merkmale fokussieren, diese Ergebnisse (vgl. [Zhe16]). Bei diesen Verfahren (Identifikationsansatz) werden im Vergleich zu den siamesischen Verfahren (Verifikationsansatz) in der Regel alle Personenlabels mitberücksichtigt, was nach Zheng et al. der Grund für die besseren Ergebnisse ist. In [Zhe16] erzielten sie bei einer Evaluation mit den Modellen *AlexNet* [Kri12], *VGG*-16 [Sim14] und *Residual*-50 [He16], die alle drei sowohl als Identifikations- sowie auch als Verifikationsansatz implementiert wurden, die jeweils besseren Ergebnisse. Die Evaluation erfolgte dabei auf dem *Market-1501*-Datensatz[3] [Zhe15]. Ein neuerer Ansatz, der nach einer Leistungsvergleichstabelle auf der *Market-Datensatz-Webseite*[4]

[3] `http://www.liangzheng.org/Project/project_reid.html`
[4] `http://www.liangzheng.org/Project/state_of_the_art_market1501.html`

die vielversprechendsten Ergebnisse liefert, ist in [Her17] vorgestellt. Mit einem *Triplet-Loss*-Ansatz für MaL auf TKFNN-Merkmalen, kombiniert mit dem Neusortierungsverfahren in [Zho17], erreichen sie die besten Ergebnisse im Vergleich zu den anderen Verfahren aus dieser Tabelle. Ein weiterer relevanter Identifikationsansatz ist in [Xia16] veröffentlicht, der auf dem CUHK1-Datensatz[5] [Li12b] den Stand der Forschung übertraf [Che17]. Die erwähnten Wiedererkennungsansätze verfolgen schwerpunktmäßig einen einzelbildbasierten Ansatz. Ein relevanter bildsequenzbasierter Ansatz ist die Verwendung von rekurrenten neuronalen Netzen, wie z.B. in [McL16].

Die in der Tabelle 2.4 aufgeführten Ansätze weisen alle sehr vielversprechende Ergebnisse auf. Mittels TKFNN gelernte Merkmale haben gegenüber den handentworfenen Merkmalen allerdings den Nachteil, dass sie für das Training eine große Menge an Bildern benötigen. Andernfalls besteht die Gefahr, dass die TKFNN schnell überangepasst werden. Folglich sollten große gelabelte Trainingsbilddaten zur Verfügung stehen und die TKFNN sollten sehr tief sein, da tiefe Netze bei gleicher Parameteranzahl besser generalisieren. Dafür sind Bildausschnitte mit hoher Auflösung nötig, um genügend Faltungs- und Gruppierungsschichten darauf anwenden zu können, was vermutlich ein Grund ist, dass TKFNN für niedrig aufgelöste Bilder nicht sehr weit verbreitet sind [Her16].

2.3 Kombinierte Personenrepräsentation

Ein naheliegender Ansatz die Schwächen der einzelnen Merkmalsarten zu kompensieren bzw. deren Stärken zu verknüpfen ist die Kombination von handentworfenen mit gelernten Merkmalen. So können beispielsweise handentworfene Merkmale, die für niedrige Auflösungen geeignet sind und eine hohe Invarianz gegenüber geometrischen Transformationen haben, mit TKFNN gelernten Merkmalen fusioniert werden, um gegenüber diesen Herausforderungen robustere Merkmale zu bekommen.

Ein TKFNN basierter Ansatz ist in [Wu17] vorgestellt, der TKFNN mit Fisher-Vektoren kombiniert, die aus SIFT-Merkmalen und Farbhistogrammen bestimmt werden. Die zugrunde liegende Anwendung ist die

[5] `http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html`

erscheinungsbasierte Personenwiedererkennung. Bei dem Ansatz werden die Fisher-Vektoren in einen Unterraum transformiert, in dem diese dann linear separiert werden können. Die Transformationen, die nichtlinear sind, werden dabei mittels einem TKFNN gelernt.

Ein weiterer vielversprechender TKFNN-Ansatz, der handentworfene Merkmale mit gelernten Merkmale fusioniert, ist in [Wu16] veröffentlicht. Dieser Fusionsansatz wurde auch im Rahmen der erscheinungsbasierten Personenwiedererkennung evaluiert, ist aber auf andere Bildauswerteaufgaben übertragbar. Die Idee hinter dem Verfahren ist, mittels einer Integration handentworfener Merkmale in den Trainingsprozess der neuronalen Netze, neue Merkmale aus einem einheitlichen Merkmalsraum zu bekommen, die diskriminativer als die jeweils einzelnen Merkmale sind. Viele erscheinungsbasierte Wiedererkennungsansätze, die einen Verifikationsansatz verfolgen (vgl. Tabelle 2.4), basieren auf einem einzigen TKFNN, das Bildpaare als Eingabe verlangt und ein Klassifikationsergebnis ausgibt, das angibt, ob die beiden Bildpaare zusammengehören. Dafür sind zahlreiche Bildpaare für jedes Anfragebild und tiefe Netze erforderlich. Im Vergleich zu diesen Ansätzen liegt in [Wu16] der Fokus auf einem Identifikationsansatz, so dass z.B. konventionelle Wiedererkennungsansätze auf den mittels TKFNN gelernten Merkmalen angewandt werden können (vgl. Abschnitt 2.2). Folglich ist dieser Ansatz auch für andere, auch nicht personenbezogene Anwendungen vielversprechend. In [Sug16] wird beispielsweise ein auf [Wu16] basierter Ansatz vorgestellt, der zu einer in einem Forum gestellten (*gepostete*) Frage aus einer riesigen Menge *geposteter* Antworten die irrelevanten herausfiltert, um die Antwort zu finden, die am *besten* die Frage beantwortet. Die Architektur zum Lernen der kombinierten Personenrepräsentation besteht aus einem *konventionellen* TKFNN, das mit einem parallelen Verarbeitungspfad für die Berechnung der handentworfenen Merkmale verknüpft ist. Beide Verarbeitungspfade verwenden denselben Bildausschnitt als Eingabe. Die kombinierte Personenrepräsentation entspricht dem Merkmalsvektor, der aus der letzten voll vernetzten Schicht extrahiert und mit dem handentworfenen Merkmal kombiniert wird. Durch Anpassung des MeL-Ansatzes *Mirror Kernel Marginal Analysis* (MKMA) aus [Che15] und Anwendung auf die TKFNN-Merkmale konnten Wu et al. auf drei öffentlichen Datensätzen für die Evaluation von Personenwiedererkennungsverfahren bessere Ergebnisse als Stand-der-Forschung-Verfahren erzielen. Dieser Ansatz wird

auch in dieser Arbeit für die Fusion von handentworfenen mit gelernten Merkmalen betrachtet (vgl. Abschnitt 8.2).

3

Personenrepräsentation mittels Kovarianzdeskriptoren

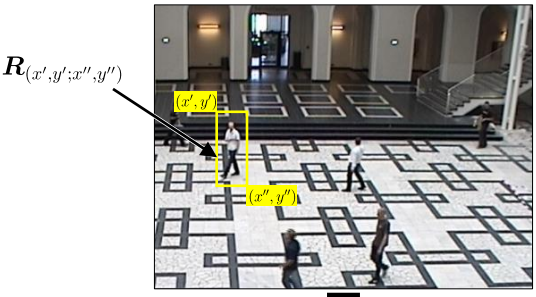
Die relevanten Arbeiten im vorherigen Kapitel verdeutlichen, dass Kovarianzdeskriptoren eine geeignete Möglichkeit darstellen, Personen in (niedrig aufgelösten) Bildern zu repräsentieren. In diesem Kapitel werden diese auf Bildregionen basierte Deskriptoren ausführlich vorgestellt und hinsichtlich den in Abschnitt 1.3 aufgeführten Herausforderungen genauer betrachtet. Kovarianzdeskriptoren können als Meta-Deskriptoren betrachtet werden, denen beliebige Bildmerkmale zugrunde gelegt werden, welche wiederum in Abhängigkeit der Anwendung oder der gewünschten Eigenschaften, wie beispielsweise Invarianzen gegenüber geometrischen Transformationen, einfach auszutauschen sind. Sie beschreiben Bildregionen mittels einer Kovarianzmatrix und werden aus einfachen Merkmalen berechnet, die innerhalb einer Bildregion liegen. Es können beliebige Merkmale dafür verwendet werden, solange diese als Skalare oder Vektoren mit reellen Zahlen vorliegen. Der Deskriptor ermöglicht somit auf eine natürliche Art und Weise mehrere Merkmale zu fusionieren, die korreliert werden können. Diagonaleinträge der Kovarianzmatrix entsprechen dabei der Varianz und die nichtdiagonalen Einträge der Korrelation zwischen den Merkmalen. Mit anderen Worten: Ein Kovarianzdeskriptor enthält Informationen über statistische und — wenn Pixelkoordinaten berücksichtigt werden — räum-

liche Eigenschaften einer Bildregion. Kovarianzmatrizen als Deskriptoren für Bildregionen einzusetzen wurde erstmals von Tuzel et al. vorgeschlagen [Tuz06]. Die Begriffe Kovarianzmatrix und Kovarianzdeskriptor werden im Folgenden synonym verwendet. Abbildung 3.1 veranschaulicht das Konzept der Kovarianzdeskriptoren.

Die Auswahl der Merkmale, die den Kovarianzdeskriptoren zugrunde gelegt werden, hängt in erster Linie von der Zielanwendung ab. Ausführliche Ergebnisse von Untersuchungen, inwiefern die Merkmalsauswahl einen Einfluss auf die Repräsentation hinsichtlich Anwendungsergebnisse haben kann, sind in [Fau15] zu finden. Unabhängig von der Merkmalsauswahl zeichnen sich Kovarianzdeskriptoren insbesondere durch ihre Stärke hinsichtlich Diskriminanz und Robustheit gegenüber Rauschen und Beleuchtungsänderungen aus [Xi13].

Eine Besonderheit gegenüber vielen anderen handentworfenen Merkmalen und Deskriptoren ist, dass die Kovarianzdeskriptoren nicht im euklidischen Raum liegen. Damit werden andere Metriken als die euklidische Metrik für die Verarbeitung von Kovarianzdeskriptoren, womit z.B. der Vergleich zweier Deskriptoren hinsichtlich *Ähnlichkeit* der zugrunde liegenden Bildregionen oder die Berechnung eines Mittelwerts aus einer Menge von Kovarianzdeskriptoren gemeint ist, notwendig. Die positiv definiten Kovarianzdeskriptoren können dafür als riemannsche Mannigfaltigkeit beschrieben werden, die im nächsten Kapitel in Abschnitt 4.1 ausführlich vorgestellt wird.

In diesem Kapitel werden zunächst allgemeine relevante Merkmals- und Deskriptortypen vorgestellt (Abschnitt 3.1) und aus Sicht der relevanten Bildauswerteverfahren betrachtet (Abschnitt 3.2). Zudem werden die wichtigsten, in dieser Dissertationsschrift einleitend beschriebenen Herausforderungen zusammenfassend als Anforderungen an Merkmale und Deskriptoren gestellt (Abschnitt 3.3). Anschließend werden in Abschnitt 3.4 die Kovarianzdeskriptoren ausführlich vorgestellt: Es wird eine effiziente Methode aus [Tuz06] zur Berechnung der Kovarianzdeskriptoren beschrieben (Abschnitt 3.4.1) und die Deskriptoren werden hinsichtlich den in dieser Arbeit einleitend aufgeführten Herausforderungen (Abschnitt 3.4.2) sowie aus Sicht der Bildauswerteverfahren bewertet (Abschnitt 3.4.3). Das wesentliche Augenmerk in dieser Arbeit liegt auf der niedrigen Auflösung,



Schritt 1:
Berechnung von Merkmalen

$$\boldsymbol{f}_{(x,y)} = \begin{pmatrix} X(x,y) \\ Y(x,y) \\ R(x,y) \\ G(x,y) \\ B(x,y) \end{pmatrix}$$

für $\boldsymbol{R}_{(x',y';x'',y'')} = \{(x,y) \mid x' \leq x < x'',\ y' \leq y < y''\}$

Schritt 2:
Berechnung des Kovarianzdeskriptors

$$\boldsymbol{\Sigma}_{\boldsymbol{R}} = \begin{pmatrix} \mathrm{Var}_{\boldsymbol{R}}(X) & \mathrm{Cov}_{\boldsymbol{R}}(X,Y) & \mathrm{Cov}_{\boldsymbol{R}}(X,R) & \mathrm{Cov}_{\boldsymbol{R}}(X,G) & \mathrm{Cov}_{\boldsymbol{R}}(X,B) \\ \mathrm{Cov}_{\boldsymbol{R}}(Y,X) & \mathrm{Var}_{\boldsymbol{R}}(Y) & \mathrm{Cov}_{\boldsymbol{R}}(Y,R) & \mathrm{Cov}_{\boldsymbol{R}}(Y,G) & \mathrm{Cov}_{\boldsymbol{R}}(Y,B) \\ \mathrm{Cov}_{\boldsymbol{R}}(R,X) & \mathrm{Cov}_{\boldsymbol{R}}(R,Y) & \mathrm{Var}_{\boldsymbol{R}}(R) & \mathrm{Cov}_{\boldsymbol{R}}(R,G) & \mathrm{Cov}_{\boldsymbol{R}}(R,B) \\ \mathrm{Cov}_{\boldsymbol{R}}(G,X) & \mathrm{Cov}_{\boldsymbol{R}}(G,Y) & \mathrm{Cov}_{\boldsymbol{R}}(G,R) & \mathrm{Var}_{\boldsymbol{R}}(G) & \mathrm{Cov}_{\boldsymbol{R}}(G,B) \\ \mathrm{Cov}_{\boldsymbol{R}}(B,X) & \mathrm{Cov}_{\boldsymbol{R}}(B,Y) & \mathrm{Cov}_{\boldsymbol{R}}(B,R) & \mathrm{Cov}_{\boldsymbol{R}}(B,G) & \mathrm{Var}_{\boldsymbol{R}}(B) \end{pmatrix}$$

Abbildung 3.1: Konzept für die Berechnung von Kovarianzdeskriptoren. Im ersten Schritt wird für jedes Pixel innerhalb der Bildregion, die durch $\boldsymbol{R}_{(x',y';x'',y'')} = \{(x,y) \mid x' \leq x < x'',\ y' \leq y < y''\}$ definiert ist, ein Merkmalsvektor $\boldsymbol{f}_{(x,y)}$ extrahiert. $\mathrm{Var}_{\boldsymbol{R}}$ und $\mathrm{Cov}_{\boldsymbol{R}}$ sind die Varianzen der einzelnen Merkmale bzw. Kovarianzen zwischen den Merkmalen für die Bildregion $\boldsymbol{R}$.

die oft für Bild- und Videoauswerteaufgaben im Videoüberwachungskontext eine der größten Herausforderungen darstellt.

3.1 Merkmale und Deskriptoren für niedrig aufgelöste Bilder von Personen

Bei niedrigen Auflösungen gibt es Einschränkungen bei der Auswahl von Merkmalen und Deskriptoren für die Repräsentation von Personen. Das soll anhand einer Gruppierung relevanter Merkmals- und Deskriptortypen in drei Gruppen — *biometrische*, *attributbasierte* und *einfache* — verdeutlicht werden.

Im Rahmen dieser Arbeit werden niedrige Auflösungen betrachtet, bei denen in der Regel keine biometrischen Merkmale von Personen extrahiert werden können. Nach der europäischen Norm *EN 50132-7* [CEN96], die u.a. für die Detektion und Wiedererkennung minimale Auflösungen von Personen in Bildern empfiehlt, sollte bei einem Bild mit 480 Zeilen die Personengröße mindestens 50% und für eine Identifizierung mindestens 120% der Bildhöhe entsprechen (vgl. auch [Mar15]). In einigen Datensätzen, die in dieser Arbeit betrachtet werden (siehe Abbildung 3.2), ist nicht einmal die Detektion von Gesichtern möglich.

Attributbasierte Ansätze erfordern in der Regel zwar auch hohe, allerdings nicht ganz so hohe Auflösungen wie biometrische Merkmale, so dass sie für niedrig aufgelöste Bilder schon eher interessant sind, insbesondere für Tracking- und Wiedererkennungsaufgaben. Sie repräsentieren Personen so ähnlich wie sie ein Mensch beschreiben würde, z.B.: *Die Person ist blond, schlank, groß, trägt ein weißes Oberteil und eine schwarze Hose sowie eine schwarze Handtasche.* Ein großer Vorteil dieser Art von Merkmalen ist, dass sie gegenüber aufnahmebedingten Störungen, kleinen Farb- und Beleuchtungsunterschieden zwischen Kameras sowie unterschiedlichen Blickwinkeln sehr robust sind (siehe z.B. [Lay12, Sch15]). Dies ermöglicht insbesondere bei der Wiedererkennung einen schnellen und zuverlässigen Abgleich von Personen. Es gibt allerdings auch einen Nachteil, der die Verwechslungsgefahr erhöht. Bei niedrig aufgelösten Bildern ist meistens

(a) **(b)** **(c)**

Abbildung 3.2: Auszüge aus Bilddatensätzen, die im Rahmen dieser Arbeit verwendet werden: Bild(er) aus dem Detektions-Datensatz **(a)**, Tracking-Datensatz **(b)** und Datensatz für die Personenwiedererkennung **(c)**. Die Bildausschnitte wurden zur Darstellung auf eine einheitliche Größe skaliert. Die Anzahl der Zeilen der Originalbilder liegt zwischen 70 und 170.

nur eine vage Personenbeschreibung möglich, da beispielsweise Handtaschen nicht detektiert werden können. Auch mögliche auffällige Muster auf einem Pullover o.ä. lassen sich dann nicht mehr aus den Bilddaten extrahieren, was die Inter-Diskiminanz deutlich senkt. Ein weiterer Nachteil attributbasierter Merkmale ist die Notwendigkeit von Trainingsdaten und einer Lernphase, in der das Videosystem bzw. der Bildauswertealgorithmus anhand von vorab annotierten Attributen trainiert werden muss.

Aus den oben genannten Gründen werden Merkmale und Deskriptoren aus der biometrischen und attributbasierten Gruppe in dieser Arbeit nicht weiter betrachtet.

Eine weitere Gruppe von Repräsentationstypen sind die einfachen Merkmale und Deskriptoren, die *nah* an der Pixelinformation sind. Sie beruhen auf lokalen Merkmalen oder globalen Ansätzen bzw. Bildregionen. Die einfachen Merkmale können in farb-, kanten-, regionen- und texturbasierte Merkmale weiter eingruppiert werden. Farbbasierte Merkmale beschreiben die Erscheinung einer Person in erster Linie anhand ihrer Kleiderfarben und werden oft durch einen Histogramm-Deskriptor repräsentiert, z.B. durch ein Farbhistogramm (z.B. [D'A11]) oder ein Histogramm mit visuellen Wörtern (z.B. [Wen11]). Kantenbasierte Ansätze konzentrieren sich auf die Kontur von Personen und werden primär für Detektions- (z.B. [Dal05]) und Trackingaufgaben (z.B. [Bil09]) verwendet, wobei sie auch schon erfolgreich

Abbildung 3.3: Repräsentation von Personen durch lokale Merkmale. Das jeweils linke Bild eines Bildpaars zeigt die Person hoch aufgelöst und das jeweils rechte Bild niedrig aufgelöst. Die linken Bilder haben im Vergleich zu den rechten Bildern neun Mal mehr Pixel.

für Wiedererkennungsverfahren eingesetzt wurden [Wan07]. Regionenbasierte Merkmale repräsentieren Objekte durch Bildregionen, die ähnliche, zusammenhängende und hauptsächlich farbige Bildpunkte zusammenfassen [For07], und die letzte Gruppe sind texturbasierte Merkmale und Deskriptoren, die neben der Farbe auch strukturelle Informationen betrachten [Hah04].

Die einfachen Merkmale und Deskriptoren sind in der Regel für niedrige Auflösungen besser geeignet. Bei den lokalen Ansätzen muss allerdings angenommen werden, dass im Allgemeinen die Anzahl lokaler Merkmale auf gering aufgelösten Bildern von Personen für eine zuverlässige Detektion anhand von *Keypoint-Detektoren* (Detektoren auf Basis lokaler Merkmale) oder für eine zuverlässige Zuordnung von Personen bei Tracking- und Wiedererkennungsaufgaben nicht ausreicht (vgl. Abbildung 3.3).

Die lokalen Merkmale wurden in dem Beispiel in Abbildung 3.3 mit dem SIFT-Algorithmus [Low04] detektiert. Das Ergebnis verdeutlicht eine Problematik bei der Auswahl von Merkmalen und Deskriptoren bei Bildern von gering aufgelösten Personen. Während die Anzahl der Merkmale bei den hoch aufgelösten Personen (jeweils das linke Bild eines Bildpaars) für eine merkmalsbasierte Personendetektion im Allgemeinen ausreicht, ist die Anzahl bei den niedrig aufgelösten Personen (jeweils das rechte

Bild eines Bildpaars) auf Grund des Detaillierungsgrads der Person zu gering für eine zuverlässige Detektion von Personen bzw. Körperteilen oder Zuordnung (Vergleich) von Personen. Merkmale und Deskriptoren, die auf Bildregionen basieren, also einen globalen Ansatz verfolgen, sind für niedrige Auflösungen oftmals besser geeignet.

3.2 Geeignete Merkmals- und Deskriptortypen aus Sicht der Bildauswerteverfahren

Dieser Abschnitt betrachtet Merkmals- und Deskriptortypen bezüglich den drei in dieser Arbeit relevanten Bildauswerteverfahren: Personendetektion, -tracking und -wiedererkennung. Aus Sicht der Bildauswerteverfahren sind Merkmale und Deskriptoren wünschenswert, die jeweils ähnlich gut bei allen drei Aufgaben abschneiden.

Detektion. Bei der Personendetektion spielen Farbinformationen eine untergeordnete Rolle, da — u.a. aufgrund der hohen Varianz — aus der Farbe hauptsächlich nur auf die Erscheinung von Personen geschlossen werden kann. Viel wichtiger ist der Personenumriss für die Detektion, anhand dessen sich sehr gut Personen in Bildern detektieren lassen. Personen können anhand einzelner Körperteile oder anhand ihres kompletten Umrisses detektiert werden.

Im Rahmen dieser Arbeit wird ein körperteilbasierter Ansatz verfolgt, weil die Vielfalt an unterschiedlichen Konturen (Intra-Diskriminanz) je Körperteil deutlich geringer ist und damit auch Personen detektiert werden können, die teilweise verdeckt sind. Dabei muss allerdings berücksichtigt werden, dass die Klassifikationsfehler im Allgemeinen zunehmen, insbesondere beim Kopf, da es viele dem Kopf ähnelnde Objektkonturen gibt. Die Anzahl der Klassifikationsfehler kann durch die Verwendung von Personenmodellen, die auf den Detektionen der Körperteile aufbauen, reduziert werden [Lei04, Jün11a]. Da der körperteilbasierte Ansatz Bilder von Wärmebildkameras verarbeiten können soll und in Wärmebildern in der Regel

die Silhouetten der Personen sehr markant sind, wurde ein konturbasiertes Detektionsverfahren erarbeitet, bei dem kantenbasierte Merkmale eingesetzt und durch Kovarianzdeskriptoren repräsentiert werden.

Tracking. Beim Tracking ist das Ziel die Verfolgung von Personen im Video, so dass im Vergleich zur Personendetektion neben den kantenbasierten Informationen auch erscheinungsbasierte Merkmale hilfreich sind, um beispielsweise Verwechslungen von Personen zu vermeiden (vgl. beispielsweise [Jep03]). Im Allgemeinen gilt, je deutlicher (auffälliger) sich eine Person anhand ihrer Erscheinung von ihrer Umgebung abhebt, desto einfacher und zuverlässiger ist sie zu tracken. Liegen Farbbilder vor, sind Farbmerkmale eine große Unterstützung. Da die direkte Verwendung von Pixelinformationen aufgrund aufnahmebedingter Störungen oder Beleuchtungsänderungen aber ungeeignet ist, sollte die erscheinungsbasierte Repräsentation durch komplexe Farbmerkmale oder passende Deskriptoren erfolgen. Viele Ansätze verwenden dafür Histogramme, um den Einfluss von Farbrauschen etc. zu mindern. In dieser Arbeit werden hierfür, wie bei der Detektion, Kovarianzdeskriptoren verwendet. Texturbasierte Merkmale sind eine weitere zusätzliche Unterstützung beim Tracking, mit denen die zu trackende Personen in bestimmten Fällen von ihrer Umgebung besser unterscheidbar wird, vorausgesetzt die Personen tragen ein texturiertes Muster. Je nach Art der Texturmerkmale können auch diese mittels Kovarianzdeskriptor mit den Farbmerkmalen fusioniert werden. Im Rahmen dieser Arbeit werden Texturmerkmale implizit durch Korrelation von Farb- mit Gradientenmerkmalen verwendet (siehe Konzept der Kovarianzdeskriptoren in Abschnitt 3.4).

Wiedererkennung. Die im Rahmen dieser Arbeit erarbeitete Personenwiedererkennung soll, wie oben motiviert, erscheinungsbasiert funktionieren. D.h. Personen sollen beispielsweise anhand ihrer Kleider- oder Haarfarbe wiedererkannt werden. Der Personenumriss spielt hierfür eine untergeordnete Rolle. Ein Abgleich von kompletten Umrissen wäre sogar von Nachteil, da die Intra-Diskriminanz bei der Wiedererkennung klein gehalten werden sollte, um Verwechslungen mit anderen Personen zu vermeiden. D.h. wiederum auch für die Auswahl von Farbmerkmalen, dass möglichst farb- und beleuchtungsinvariante Merkmale extrahiert bzw. diese durch einen

geeigneten Deskriptor invariant bzgl. solcher Unterschiede gemacht werden sollten. Darüber hinaus sollte die Personenrepräsentation auch invariant gegenüber unterschiedlichen Blickwinkeln sein, um die Intra-Diskriminanz klein zu halten. Die Inter-Diskriminanz sollte dagegen erhöht werden, beispielsweise durch Verwendung texturbasierter Merkmale.

3.3 Anforderungen an handentworfene Merkmale und Deskriptoren

Neben der niedrigen Bild- und Videoauflösung gibt es noch weitere Herausforderungen, die bei der Auswahl und dem Entwurf von Merkmalen und Deskriptoren berücksichtigt werden sollten. Merkmale und Deskriptoren für die in dieser Arbeit relevanten Bildauswerteaufgaben sollten robust oder invariant gegenüber Blickwinkel- und Beleuchtungsänderungen sowie Veränderungen der Personenkontur sein. Zudem sollten sie mit Verdeckungen, Schatten, schwachem Kontrast der Personen zum Hintergrund und generellen Hintergrundstörungen umgehen können sowie auch mit aufnahmebedingten Störungen, wie z.B. Kompressionsartefakten, niedrigen Bildwiederholraten, Zeilensprüngen, Unschärfe, Bild- und Farbrauschen (vgl. Abschnitt 1.3).

Werden maschinelle Lernverfahren auf die Merkmale und Deskriptoren angewandt, können die Probleme teilweise auch durch diese Verfahren gelöst werden [Bis09]. Merkmale können z.B. (überwacht) gelernt (vgl. Abschnitt 2.2) und handentworfene Merkmale mittels MeL oder Dimensionsreduktionsverfahren optimiert werden (vgl. Abschnitt 2.1.5), um hinsichtlich den Herausforderungen robuster zu werden. Der Fokus bei den handentworfenen Ansätzen sollte allerdings auf dem Entwurf der Merkmale und Deskriptoren liegen, wie durch das folgende Beispiel der erscheinungsbasierten Personenwiedererkennung deutlich wird.

Anwendungsbeispiel Personenwiedererkennung. Soll die erscheinungsbasierte Personenwiedererkennung unüberwacht mit handentworfenen Merkmalen durchgeführt werden, ist die Auswahl bzw. Konstruktion geeigneter Merkmale und Deskriptoren ein wichtiger Schritt. Erschwerend

Abbildung 3.4: Beispielbildausschnitte segmentierter Personen, die mittels einem bewegungsbasierten Segmentierungsverfahren in komprimierten Videoüberwachungsszenarien vom Hintergrund getrennt wurden.

kommt bei dieser Aufgabe hinzu, dass die Personen möglichst exakt segmentiert werden sollten. Bei Deskriptoren für Bildregionen beispielsweise sollte die Bildregion so um eine Person gelegt werden, dass die Person zentriert und vollständig, aber mit möglichst wenig Hintergrund durch die Bildregion erfasst wird. Eine exakte Segmentierung der Person ist dabei von großem Nutzen, die jedoch aufgrund von Hintergrundstrukturen, Kompressionsartefakten etc. in der Regel schwierig zu erstellen ist (vgl. Abbildung 3.4).

Wie durch die Abbildung 3.4 exemplarisch gezeigt wird, stehen meist lediglich grobe Segmentierungen zur Verfügung, so dass auch Hintergrundinformationen in die Personenrepräsentation mit einfließen, was die Wiedererkennung beeinträchtigen kann. Oft werden die Kameradaten jedoch überhaupt nicht vorprozessiert, weshalb die Personenwiedererkennungsverfahren gesamte Bildausschnitte mit unsegmentierten Personen verarbeiten müssen. Eine weitere Schwierigkeit bei dieser Aufgabe sind eventuell vorhandene Farbunterschiede, die in Kameranetzwerken aufgrund unterschiedlicher Konfigurationen, Beleuchtungsverhältnisse oder Kameramodelle zwischen den einzelnen Kameras auftreten können. Da die Wiedererkennung von Personen in niedrig aufgelösten Bildern meist nur erscheinungsbasiert durchgeführt werden kann, also anhand der Kleidung, sollten die Merkmale und Deskriptoren robust gegenüber solchen Farbunterschieden sein. Eine Alternative wäre die Bestimmung von Transferfunktionen, welche die Helligkeits- und Farbunterschiede zwischen den Kameras ausgleichen [Ily10]. Bei überlappenden Sichtfeldern von Kameras kann dies automatisch erfolgen. Überlappen sich die Sichtfelder allerdings nicht, müssen die installierten Kameras zunächst z.B. anhand eines

Farbmusters aufeinander abgestimmt werden, was den Aufwand bei der Inbetriebnahme und dem Austausch von Kameras in Kameranetzwerken deutlich erhöht. Darüber hinaus sollten die Merkmale und Deskriptoren — was auch für Detektions- und Trackingverfahren wichtig ist — unter den oben genannten Anforderungen diskriminativ sein. Die Ähnlichkeit von Deskriptoren unterschiedlicher Personen sollte niedrig (Inter-Diskriminanz) und von Deskriptoren derselben Person hoch (Intra-Diskriminanz) sein.

Im Folgenden werden die Anforderungen aufgelistet, die an handentworfene Merkmale und Deskriptoren und damit auch an die Kovarianzdeskriptoren gestellt werden. Die Anforderungen ergeben sich aus den Herausforderungen in Abschnitt 1.3 und gelten nicht nur für die Personenwiedererkennungsansätze, sondern berücksichtigen auch Detektions- und Trackingverfahren.

Anforderungen an die Kovarianzdeskriptoren:

- Da das Hauptaugenmerk in dieser Arbeit auf der niedrigen Auflösung liegt, sollten die Kovarianzdeskriptoren dafür entsprechend geeignet sein.

- Insbesondere bei kameraübergreifenden Anwendungen stellen Beleuchtungsänderungen und Farbunterschiede zwischen Kameras, die sowohl aufnahmebedingte als auch anwendungsbezogene Gründe haben können, große Herausforderungen dar, die während der Deskriptorberechnung begegnet werden sollten.

- Weitere aufnahmebedingte Störungen wie z.B. Rauschen, Zeilensprünge, Kompressionsartefakte, Unschärfe, etc. sollten auch beim Entwurf von Deskriptoren und nicht erst in der folgenden Verarbeitungskette berücksichtigt werden.

- Je nach Anwendung können invariante Eigenschaften einen großen Vorteil bieten. Die Deskriptoren sollten prinzipiell so entworfen werden können, dass sie invariant gegenüber unterschiedliche Ansichten, Orientierungen und Größen der Personen sind. Andernfalls müssen solche Variationen in der weiteren Verfahrenskette explizit behandelt werden.

- Außerdem wäre die Berücksichtigung von anwendungsbezogenen Herausforderungen vorteilhaft, wie z.B. eine diskriminative Repräsentationen von Personen ähnlicher Erscheinung.

3.4 Kovarianzdeskriptoren

Bevor die Kovarianzdeskriptoren hinsichtlich der für diese Arbeit relevanten Bildauswerteverfahren und Herausforderungen diskutiert werden, wird die effiziente Methode aus [Tuz06] zur Berechnung der Kovarianzdeskriptoren beschrieben.

3.4.1 Berechnung

In diesem Abschnitt wird eine Methode zur effizienten Berechnung von Kovarianzdeskriptoren aufgeführt, wie sie in [Tuz06, Por06a] vorgeschlagen wird. Die Berechnung basiert auf Integralbildern, die zur schnellen Berechnung von Mittelwertbildern aus Bildstapeln verwendet werden [Cro84]. Mit Hilfe von Integralbildern lassen sich Pixelwertsummen innerhalb Bilder bzw. rechteckiger Bildausschnitte in konstanter Zeit berechnen. Jeder Pixelwert eines Integralbilds entspricht der Summe aller Pixelwerte innerhalb des rechteckigen Bildausschnitts, die durch die Position des Pixels und der oberen linken Ecke des rechteckigen Bildausschnitts definiert ist. Sei $\boldsymbol{I_R}$ ein Bildausschnitt eines Intensitätsbilds, das durch die obere linke Ecke $(0,0)$ und die untere rechte Ecke (x'', y'') definiert ist (vgl. Abbildung 3.5). Das Integralbild $\boldsymbol{J_R}$ von $\boldsymbol{I_R}$ ist dann gegeben durch

$$\boldsymbol{J}(x'', y'') = \sum_{x<x'', y<y''} \boldsymbol{I}(x, y)\,, \quad x, y \in \boldsymbol{R}_{(0,0;x'',y'')}\,. \tag{3.1}$$

Die Berechnung von Kovarianzdeskriptoren erfolgt gemäß [Tuz06]. Sei $\boldsymbol{R}_{(x',y';x'',y'')}$ ein rechteckiger Bildausschnitt in einem RGB-Bild mit der Breite b und Höhe h, der auf der Pixelmenge $\{(x, y) \mid x' \leq x < x'',\ y' \leq y < y''\}$ definiert ist. Im ersten Schritt werden für jedes Pixel (x, y) aus dieser Menge Merkmale berechnet. Es können beliebige Merkmale verwendet werden, vorausgesetzt die Merkmale können durch Skalare

oder Vektoren repräsentiert werden. Sollen beispielsweise die Koordinaten ($x = X(x,y), y = Y(x,y)$) sowie Rot-, Grün- und Blauwerte der Pixel ($R(x,y)$,$G(x,y)$,$B(x,y)$) zur Repräsentation des Bildausschnitts $\boldsymbol{R}_{(x',y';x'',y'')}$ verwendet werden, wird für jedes Pixel des Bildausschnitts $\boldsymbol{R}_{(x',y';x'',y'')}$ ein 5-dimensionaler Merkmalsvektor $\boldsymbol{f}_{(x,y)}$ bestimmt:

$$\boldsymbol{f}_{(x,y)} = \begin{pmatrix} X(x,y) \\ Y(x,y) \\ R(x,y) \\ G(x,y) \\ B(x,y) \end{pmatrix} . \tag{3.2}$$

Die Varianz der Pixelpositionen ist zwar für alle Bildregionen der gleichen Größe identisch, dennoch ist deren Verwendung wichtig, um räumliche Informationen zu erhalten. Die Korrelation der Intensität mit den Pixelpositionen beispielsweise gibt Rückschluss über die räumliche Verteilung der Intensitätswerte von Pixeln.

Der Kovarianzdeskriptor $\boldsymbol{\Sigma}_{\boldsymbol{R}}$ für die Bildregion $\boldsymbol{R}_{(0,0;b,h)}$ ergibt sich aus der Gleichung

$$\boldsymbol{\Sigma}_{\boldsymbol{R}} = \frac{1}{b \cdot h} \sum_{x=0}^{b-1} \sum_{y=0}^{h-1} \left(\boldsymbol{f}_{(x,y)} - \boldsymbol{\mu}_{\boldsymbol{R}} \right) \left(\boldsymbol{f}_{(x,y)} - \boldsymbol{\mu}_{\boldsymbol{R}} \right)^T , \tag{3.3}$$

wobei $\boldsymbol{\mu}_{\boldsymbol{R}}$ der Mittelwertvektor der Merkmalsvektoren $\{\boldsymbol{f}_{(x,y)}\}$ ist ($x = 0, \ldots, b-1,\ y = 0, \ldots, h-1$), der gegeben ist durch

$$\boldsymbol{\mu}_{\boldsymbol{R}} = \frac{1}{b \cdot h} \sum_{x=0}^{b-1} \sum_{y=0}^{h-1} \boldsymbol{f}_{(x,y)} . \tag{3.4}$$

Eine schnelle Berechnung von Histogrammen für Bildregionen mittels Integralbildern ist in [Por05] vorgeschlagen und in [Tuz06, Por06a] in einer angepassten Version für die Berechnung von Kovarianzdeskriptoren aufgeführt. Der angepasste Ansatz wird im Folgenden gemäß [Tuz06, Por06a] zusammengefasst.

Insgesamt müssen $e + e^2$ Integralbilder berechnet werden, wobei e der Dimension der Merkmalsvektoren entspricht. Zur effizienten Berechenbarkeit

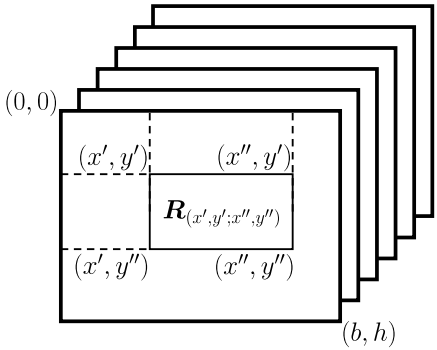


Abbildung 3.5: Veranschaulichung des 3D-Arrays $F[\cdot,\cdot,\cdot]$ (Array von Merkmalsvektoren $\boldsymbol{f}_{(x,y)}$) und Definition des Rechtecks $\boldsymbol{R}$. Die einzelnen dargestellten Ebenen entsprechen $F[\cdot,\cdot,a]$, $a = 0,\ldots,e-1$. Das Rechteck $\boldsymbol{R}$ ist durch $\boldsymbol{R}_{(x',y';x'',y'')} = \{(x,y) \mid x' \le x < x'',\, y' \le y < y''\}$ definiert.

der Kovarianzdeskriptoren werden höherdimensionale Arrays im Speicher angelegt.

Sei $F[\cdot,\cdot,\cdot]$ ein 3D-Array von Merkmalsvektoren, das durch folgende Gleichung definiert ist:

$$F[x,y,a] := f_{(x,y)}(a) \tag{3.5}$$

wobei a die Merkmalsdimension angibt und $\boldsymbol{f}_{(x,y)}$ dem Merkmalsvektor für den Pixel mit der Koordinate (x,y) entspricht. Sei weiter $P[\cdot,\cdot,\cdot]$ ein 3D-Array von Integralbildern für die einzelnen Merkmalsdimensionen $a = 0,\ldots,e-1$, das definiert ist durch

$$P[x',y',a] := \sum_{x<x',y<y'} F[x,y,a], \quad a = 0,\ldots,e-1 \tag{3.6}$$

und $Q[\cdot,\cdot,\cdot,\cdot]$ ein 4D-Array von Integralbildern für die einzelnen Merkmalsdimensionen $a,b = 0,\ldots,e-1$, das definiert ist durch

$$Q[x',y',a,b] := \sum_{x<x',y<y'} F[x,y,a]F[x,y,b], \quad a,b = 0,\ldots,e-1\,. \tag{3.7}$$

Der Kovarianzdeskriptor für den rechteckigen Bildausschnitt $\boldsymbol{R}_{(0,0;x',y')}$ ist dann gegeben durch

$$\boldsymbol{\Sigma}_{\boldsymbol{R}_{(0,0;x',y')}} = \frac{1}{n_0 - 1}\left(\boldsymbol{Q}_{x',y'} - \frac{1}{n_0}\,\boldsymbol{p}_{x',y'}\,\boldsymbol{p}_{x',y'}^T\right), \tag{3.8}$$

mit $n_0 := x'y'$ und

$$\begin{aligned}\boldsymbol{p}_{x,y} &:= (P[x,y,0]\ldots P[x,y,e-1])^T,\\ \boldsymbol{Q}_{x,y} &:= \begin{pmatrix} Q[x,y,0,0] & \ldots & Q[x,y,0,e-1]\\ \vdots & \ddots & \vdots\\ Q[x,y,e-1,0] & \ldots & Q[x,y,e-1,e-1]\end{pmatrix}.\end{aligned} \tag{3.9}$$

Durch Umformen ergibt sich folgende Gleichung für die Berechnung einer beliebigen rechteckigen Bildregion $\boldsymbol{R}_{(x',y';x'',y'')} = \{(x,y) \mid x' \leq x < x'',\ y' \leq y < y''\}$, wie er in Abbildung 3.5 dargestellt ist:

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{R}_{(x',y';x'',y'')}} =& \frac{1}{n_1 - 1}\Bigg(\boldsymbol{Q}_{x'',y''} + \boldsymbol{Q}_{x',y'} - \boldsymbol{Q}_{x'',y'} - \boldsymbol{Q}_{x',y''}\\ &- \frac{1}{n_1}\left(\boldsymbol{p}_{x'',y''} + \boldsymbol{p}_{x',y'} - \boldsymbol{p}_{x',y''} - \boldsymbol{p}_{x'',y'}\right)\\ &\left(\boldsymbol{p}_{x'',y''} + \boldsymbol{p}_{x',y'} - \boldsymbol{p}_{x',y''} - \boldsymbol{p}_{x'',y'}\right)^T\Bigg),\end{aligned} \tag{3.10}$$

mit $n_1 := (x'' - x')(y'' - y')$.

3.4.2 Eigenschaften

Kovarianzdeskriptoren eignen sich sehr gut für die Repräsentation von Bildern gering aufgelöster Personen, da diese Deskriptoren auch Bildausschnitte repräsentieren können, für die nur wenige Merkmale extrahiert werden können (vgl. Abschnitt 3.1). Mittels Beschreibung einer Bildregion durch statistische Eigenschaften zweiter Ordnung steigt zudem — gegenüber der Betrachtung der absoluten Pixelwerte — die *Robustheit* der Personenrepräsentation. Durch die Verwendung von Kovarianzdeskriptoren anstatt die zugrunde liegenden Merkmale direkt zu verwenden, können in Bildauswerteaufgaben höhere Detektionsraten erzielt werden, wie z.B. in

[Tuz07] und [Pai07] gezeigt wurde. Ein Grund dafür kann beispielsweise die implizite Rauschreduktion bei der Berechnung der Deskriptoren oder die Invarianz der Deskriptoren gegenüber linearen Verschiebungen von Merkmalen sein. Im Rahmen der Anwendung *Personendetektion anhand von Körperteilen in Einzelbildern* in Kapitel 5 wurde ein für Konturen von Körperteilen angepasstes Detektionsverfahren, das auf der Arbeit [Dal05] basiert, als Referenzverfahren erarbeitet (siehe Abschnitt 5.1) und mit dem Verfahren für die Körperteildetektion (siehe [Met09] und Abschnitt 5.2) verglichen, das die obige Aussage bekräftigt.

Bei hoch aufgelösten Bildern von Personen hingegen ist ein einzelner Kovarianzdeskriptor in den meisten Fällen allerdings ungeeignet, da der Deskriptor beispielsweise hoch aufgelöste Texturen nicht detailreich repräsentieren kann. Dazu müsste der Bildausschnitt zunächst in mehrere kleinere Bereiche aufgeteilt, für jeden ein Deskriptor berechnet und daraus ein Modell bestimmt werden. Ein solcher Ansatz ist z.B. in [Bak10] veröffentlicht. Dabei wird der Bildausschnitt anhand von Körperpartien, die mittels eines Detektors automatisch bestimmt werden, weiter unterteilt. Einfachere Ansätze teilen die Bildausschnitte in horizontale Streifen, wobei die Anzahl und Höhe der Steifen meist fest vorgegeben wird. Ein solcher Ansatz wird beispielsweise in [Lia15] verfolgt, der allerdings auf LOMO-Merkmalen basiert, aber auch für Kovarianzdeskriptoren geeignet ist (siehe z.B. [Hir11]). Durch die Betrachtung einzelner Streifen wird dabei die Invarianz gegenüber dem Blickwinkel bzw. der (horizontalen) Orientierung der Person erhöht, jedoch eventuell unter Einbußen der Diskriminanz [Lia15]. Im Rahmen dieser Arbeit werden solche hierarchischen Modelle nicht behandelt. Aufgrund der betrachteten Auflösungen wird angenommen, dass ein Objekt immer geeignet durch genau einen Kovarianzdeskriptor repräsentiert werden kann.

Ein weiterer Vorteil der Betrachtung von Varianzen und Kovarianzen anstatt den Pixelwerten bzw. Merkmalen direkt ist die Robustheit gegenüber Rauschen. Abbildung 3.6 illustriert dies anhand eines Beispiels. Während der Berechnung der Kovarianzmatrix wird das Rauschen aufgrund der Mittelwertberechnung reduziert. Diese Mittelwert-Filterung ist auch bei anderen aufnahmebedingten Störungen wie z.B. Kompressionsartefakten, Linsenverzerrungen oder Unschärfe vorteilhaft.

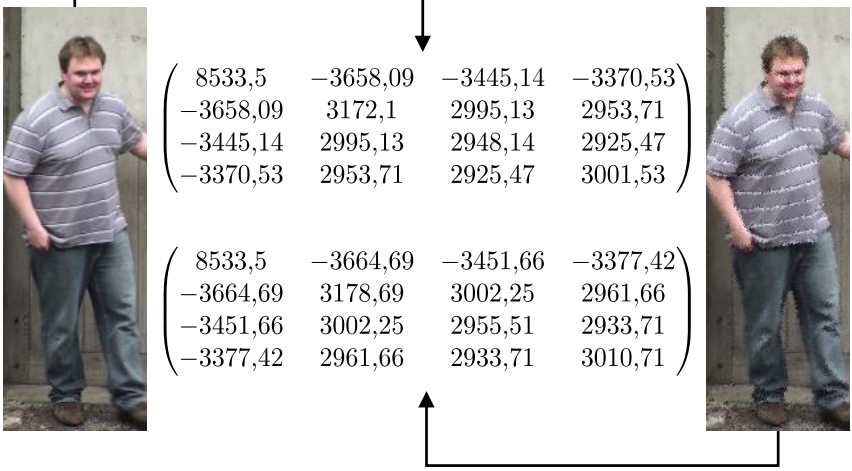


Abbildung 3.6: Kovarianzdeskriptoren für zwei Bildausschnitte, die bis auf einen unterschiedlichen Rauschanteil identisch sind. Dem rechten Bild wurde vor der Berechnung des Deskriptors ein künstliches Rauschen hinzugefügt. Die Kovarianzdeskriptoren wurden aus y-Koordinaten und Farbwerten (RGB-Farbraum) berechnet. Die Zeilen der Matrizen von oben nach unten bzw. deren Spalten von links nach rechts entsprechen der folgenden Reihenfolge der einfachen Merkmale: y-Koordinaten, Rot-, Grün-, Blauwerte. Die Varianzen und Kovarianzen unterscheiden sich nur gering zwischen den beiden Matrizen, was sich im Abstand widerspiegelt, der $4,9 \cdot 10^{-15}$ ist (die hier verwendete Abstandsfunktion ist in Gleichung (4.5) definiert).

Eine Repräsentation einer Person durch ausschließlich Varianzen und Kovarianzen erzeugt identische Deskriptoren bei gleicher Verteilung und unterschiedlichem Mittelwert von Merkmalen. Dadurch wird die Invarianz gegenüber linearen Verschiebungen von Merkmalen erhöht, so dass Kovarianzdeskriptoren z.B. gut mit globalen Helligkeits- und Farbunterschieden zurechtkommen, die zwischen Bildern unterschiedlicher Kameras auftreten können. Dies wird durch die Abbildung 3.7 beispielhaft illustriert.

Eine Robustheit von Kovarianzdeskriptoren gegenüber Helligkeits- und Farbunterschieden ist bekannt [Ma12a, Xi13]. Dies ist ein großer Vorteil für den Vergleich von Bildregionen in Kameranetzwerken, sowohl hinsichtlich der Güte des Vergleichsergebnisses als auch in punkto Praxiseinsatz. So können beispielsweise Objektwiedererkennungsverfahren, die auf Kovarianzdeskriptoren basieren, in bestehenden Kameranetzwerken eingesetzt werden, ohne dass in der Regel eine Farbkalibrierung der Kameras durchgeführt werden muss.

Hinsichtlich Verdeckungen von Personen sind Kovarianzdeskriptoren aufgrund ihres globalen Charakters weniger gut geeignet. Ansätze mit lokalen Merkmalen können in der Regel besser mit Verdeckungen umgehen. Dabei spielt es keine Rolle, ob die Personen durch andere Personen oder Objekte verdeckt werden. Bei Wiedererkennungsaufgaben zwischen verschiedenen Kameras können folglich sogar Verschmutzungen auf Kameraobjektiven etc. eine Herausforderung darstellen. Diese Herausforderungen sind dann durch weitere Bildauswerteverfahren in der Verarbeitungskette zu lösen.

Für *erscheinungsergänzende* Effekte wie beispielsweise Schatten, die in der Regel eine Beeinträchtigung darstellen, können globale Repräsentationen gegenüber lokalen Ansätzen jedoch vorteilhaft sein. Während im Allgemeinen bei lokalen Repräsentationsansätzen die Gefahr besteht, dass lokale Merkmale — die beispielsweise auf die stärksten Kanten springen — auf den Schatten *überspringen*, also die Person überwiegend durch den Schatten repräsentieren, stellen bei globalen Ansätzen Schatten nur eine *Ergänzung* der Repräsentation dar.

Kovarianzdeskriptoren können außerdem so konstruiert werden, dass sie invariant gegenüber geometrischen Transformationen sind. Ob und gegenüber welchen Transformationen sie invariant sind, ist von der Auswahl der Merkmale abhängig. So ist ein Deskriptor nicht rotationsinvariant, wenn

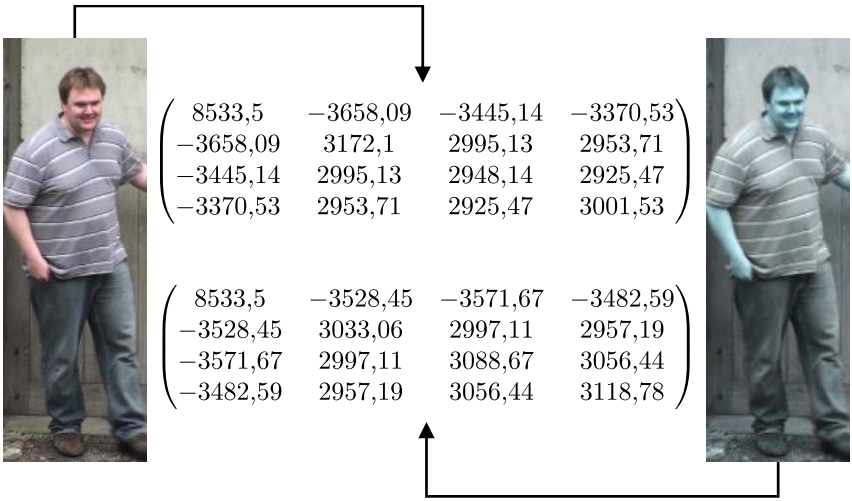


Abbildung 3.7: Kovarianzdeskriptoren für zwei Bildausschnitte, die bis auf einen unterschiedlichen Farbton identisch sind. Die Kovarianzdeskriptoren wurden wieder aus y-Koordinaten und Farbwerten (RGB-Farbraum) berechnet und die Reihenfolge entspricht Abbildung 3.6. Der Farbtonunterschied erzeugt einen Abstand von $0,39$, der zwar höher als bei dem Beispiel mit dem künstlichen Rauschen in Abbildung 3.6 ist, aus praktischer Sicht jedoch nach wie vor als klein angesehen werden kann.

beispielsweise neben der Hauptgradientenrichtung auch die Bildkoordinaten als Merkmale verwendet werden. Bei Rotationen eines Bildausschnitts, die insbesondere bei Luftbildauswertungen relevant sind, würden sich dann in den meisten Fällen die Kovarianzen zwischen den Koordinaten und den Gradientenrichtungen ändern. Wird die Hauptgradientenrichtung allerdings bezüglich einer vorgegebenen Achse angegeben — in der Luftbildauswertung könnte das z.B. eine Symmetrieachse sein, die durch die Schultern der Personen läuft — wird die Repräsentation robust gegenüber Rotationen. Ein Ansatz, der solch eine ähnliche Strategie verfolgt, ist z.B. SDALF, wobei die Personenrepräsentation bzgl. der vertikalen Symmetrieachse von Personen aufgebaut wird (bei Seitenansichten). Die Achse bzw. die Pose der Personen wird dabei automatisch geschätzt. Eine ausführliche Untersuchung, was für einen Einfluss die Merkmalsauswahl und Bildtransformationen haben, ist in [Fau15] gegeben.

3.4.3 Fazit aus Sicht der Bildauswerteverfahren

Aufgrund der oben genannten Faktoren und Überlegungen sowie der Eigenschaft der Kovarianzdeskriptoren einfache Merkmale effizient fusionieren zu können, ist deren Einsatz aus Sicht der Bildauswerteverfahren sinnvoll. Kovarianzdeskriptoren repräsentieren eine Personen dabei nicht durch die absoluten Werte der extrahierten Merkmale, sondern durch deren Varianzen und Korrelationen untereinander. Mit dieser *Meta-Beschreibung* von Merkmalen stellt diese Repräsentation damit eine *Brücke* zwischen einer Repräsentation durch einfache und einer Beschreibung durch attributbasierte Merkmale dar.

Das erste Bild (Bildausschnitt) des Bildtripels 3.2 (c) (Person mit grünem Oberteil) ist ein gutes Beispiel für die Illustration der *semantischen* Repräsentation von Merkmalen durch Kovarianzdeskriptoren, wenn man z.B. die Grünwerte der Pixel und die Korrelation dieser Werte mit den y-Koordinaten betrachtet. Der Kovarianzdeskriptor beschreibt die Person anhand der Varianz des Grünwerts, die in diesem Beispiel auf einen hohen Grünanteil hindeutet, sowie durch den Korrelationswert, der angibt, dass sich der Grünanteil im oberen Bildbereich befindet.

Aus Sicht der Bildauswerteverfahren unterscheiden sich Kovarianzdeskriptoren bezüglich Anwendung einzig in der Zusammensetzung der einfachen

Merkmale, wobei auch eine generalisierte Personenrepräsentation für alle drei unterschiedlichen Bildauswerteaufgaben definiert werden kann. Dazu muss eine Merkmalskombination gewählt werden, die für alle drei Bildauswerteaufgaben geeignet ist. In [Fau15] wird untersucht, welchen Einfluss die Merkmalsauswahl auf die Anwendungsergebnisse haben können.

3.4.4 Zusammenfassung

Zusammenfassend betrachtet bieten Kovarianzdeskriptoren einige Vorteile, weshalb sie im Rahmen dieser Arbeit bei allen drei Bildauswerteaufgaben eingesetzt werden. Durch ihren generischen Charakter — die Repräsentation von Personen durch Bildausschnitte — sind sie prädestiniert für niedrige Auflösungen. Durch die Repräsentation mittels Varianzen und Kovarianzen von bzw. zwischen Merkmalen anstelle durch Merkmale direkt, sind die Kovarianzdeskriptoren wegen der impliziten Mittelwert-Filterung zudem robust gegenüber Rauschen, Beleuchtungs- und Farbänderungen. Auch der Einfluss von Kompressionsartefakten, Unschärfe, Zeilensprünge, Farbsäume etc. wird gemindert. Viele der in Abschnitt 1.3 angesprochenen Herausforderungen werden damit durch die gewählte Personenrepräsentation begegnet.

4

Mathematisches Rahmenwerk zur Anwendung von Kovarianzdeskriptoren

In diesem Kapitel wird das im Rahmen dieser Arbeit verwendete mathematische Rahmenwerk zur Verarbeitung und Anwendung von positiv definiten Kovarianzdeskriptoren vorgestellt. Es basiert auf den Arbeiten in [Pen06b], dem darauf aufbauenden riemannschen Rahmenwerk für Tensorberechnung in [Pen06a] und den in [Pen99] vorgestellten Wahrscheinlichkeitsmethoden für riemannsche Mannigfaltigkeiten.

Darüber hinaus werden im letzten Teil dieses Kapitels nichtlineare Dimensionsreduktionsverfahren behandelt, die einen Teil des Rahmenwerks bilden. Der Fokus liegt dabei auf dem MaL Algorithmus *Laplacian Eigenmaps* (LE) [Bel03], der im Rahmen dieser Arbeit eingesetzt wird. Der Algorithmus wird auf der riemannschen Mannigfaltigkeit der Kovarianzdeskriptoren angewandt, mit dem Ziel, die in dieser Arbeit erarbeiteten Verfahren zur Personendetektion und -wiedererkennung zu verbessern. Die Erweiterung dieser Verfahren mittels MaL wird im Anschluss an das mathematische Rahmenwerk in den jeweiligen Kapiteln der Personendetektion und -wiedererkennung vorgestellt.

4.1 Riemannsche Mannigfaltigkeit der Kovarianzdeskriptoren

Damit der Kovarianzdeskriptor zur Detektion, Verfolgung und Wiedererkennung von Personen eingesetzt werden kann, benötigt man eine geeignete Metrik und Operationen, um beispielsweise die Ähnlichkeit zwischen den symmetrischen und positiv definiten Kovarianzdeskriptoren bestimmen zu können. Die Bestimmung der Ähnlichkeiten anhand herkömmlich definierter Matrixoperationen wie z.B. der Spur (weil dabei die Kovarianzen nicht berücksichtigt werden) oder der Determinante, wie man anhand folgendem Beispiel einfach erkennen kann, ist ungeeignet:

$$\det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \det \begin{pmatrix} 0,02 & 0 \\ 0 & 50 \end{pmatrix} . \tag{4.1}$$

Trotz der großen Differenzen zwischen den Varianzen der einzelnen Matrizen ist die Determinante gleich. Außerdem sind die herkömmlich definierten Operationen eines reellen Vektorraums oder die gewöhnlichen Matrixoperationen (für symmetrische Matrizen) wie z.B. das Matrixexponential oder der Logarithmus einer Matrix für die Verarbeitung von Kovarianzdeskriptoren ungeeignet, weil der Raum der Kovarianzdeskriptoren kein Matrizenraum mit der gewöhnlichen additiven Struktur ist. Die Operationen können Matrizen erzeugen, die beispielsweise negative Eigenwerte besitzen und damit z.B. die Bedingung der positiven Definitheit von positiv definiten Kovarianzmatrizen verletzen, was wiederum dazu führen kann, dass im Vektorraum ähnliche Kovarianzdeskriptoren einen großen Abstand und unähnliche einen kleinen Abstand zueinander haben. Auch bei anderen Aufgaben, wie beispielsweise der Interpolation von Kovarianzmatrizen, erweisen sich euklidische Ansätze als ungeeignet (siehe z.B. [Ars06]). Im Folgenden wird ein geeignetes mathematisches Rahmenwerk zur Verarbeitung und zum Vergleich von Kovarianzdeskriptoren vorgestellt, das u.a. die Bedingung der positiven Definitheit einhält.

Der Vektorraum der $n \times n$ Matrizen ist ein n^2-dimensionaler Matrizenraum, worin die symmetrischen $n \times n$ Matrizen einen $\frac{n \cdot (n+1)}{2}$-dimensionalen Unterraum bilden. Die positiv definiten $n \times n$ Kovarianzdeskriptoren beschreiben im Unterraum der symmetrischen $n \times n$ Matrizen einen offenen Halbkegel mit derselben Dimension (vgl. [Pen06b]), der nicht abgeschlossen

unter Vektoraddition ist. Die gewöhnlichen Matrixoperationen, wie z.B. die Subtraktion zweier Matrizen, können Matrizen erzeugen, die beispielsweise negative Eigenwerte besitzen und damit keinem Kovarianzdeskriptor mehr entsprechen. Um das zu vermeiden, werden Mannigfaltigkeiten — topologische Räume, die lokal dem euklidischen Raum gleichen — betrachtet. In [Pen06a] wird der Halbkegel (flache unvollständige Mannigfaltigkeit) durch eine reguläre und vollständige — allerdings gekrümmte — riemannsche Mannigfaltigkeit ersetzt, die im Rahmen dieser Arbeit betrachtet wird. Diese riemannsche Mannigfaltigkeit von positiv definiten $n \times n$ Kovarianzdeskriptoren wird im Folgenden als Raum der Kovarianzdeskriptoren Sym_n^+ bezeichnet. Zudem wird die darin vorgeschlagene affin-invariante riemannsche Metrik in dieser Arbeit verwendet. In dem Raum Sym_n^+ gibt es immer eine — und nur eine — Geodätische zwischen zwei Kovarianzdeskriptoren, so dass die Geodäte global — und nicht nur lokal — die kürzeste Verbindungskurve zweier Kovarianzdeskriptoren ist. Die Länge der Geodätischen entspricht somit dem (geodätischen) Abstand von zwei Kovarianzdeskriptoren im Sym_n^+. Bevor im Folgenden auf die riemannsche Mannigfaltigkeit Sym_n^+ ausführlicher eingegangen wird, wird zunächst das allgemeine Konzept der riemannschen Mannigfaltigkeit kurz dargestellt.

Eine riemannsche Mannigfaltigkeit (M, g) ist eine differenzierbare n-dimensionale Mannigfaltigkeit M, die lokal einem euklidischen Raum $\mathbb{R}^n$ gleicht (global muss M dem $\mathbb{R}^n$ nicht ähneln). Da M differenzierbar ist, gibt es an jedem Punkt $\boldsymbol{P} \in M$ einen Tangentialraum (einen Vektorraum $T_{\boldsymbol{P}}M$, der M (lokal) am Punkt $\boldsymbol{P}$ bis zur ersten Ordnung approximiert). Außerdem ist eine riemannsche Mannigfaltigkeit mit einer riemannschen Metrik g versehen, die jedem Punkt $\boldsymbol{P} \in M$ ein Skalarprodukt $\langle\,|\,\rangle_{\boldsymbol{P}} : T_{\boldsymbol{P}}M \times T_{\boldsymbol{P}}M \to \mathbb{R}$ zuordnet, mit dem eine Abstandsfunktion auf M definiert werden kann. Der geodätische Abstand $\mathrm{g}(\boldsymbol{P}_1, \boldsymbol{P}_2)$ zwischen zwei Punkten $\boldsymbol{P}_1 \in M$ und $\boldsymbol{P}_2 \in M$ entspricht der Länge (Bogenlänge) einer kürzesten Verbindungskurve, die $\boldsymbol{P}_1$ und $\boldsymbol{P}_2$ verbindet. Er kann aufgrund der lokalen Ähnlichkeit von M zu einem euklidischen Raum durch das Längenfunktional bestimmt werden, das lokal definiert ist durch

$$L(f) = \int_0^1 \sqrt{\langle \frac{\mathrm{d}}{\mathrm{d}\,t} f(t) \,|\, \frac{\mathrm{d}}{\mathrm{d}\,t} f(t) \rangle}\, \mathrm{d}\,t\,, \tag{4.2}$$

wobei $f : [0, 1] \to M$ eine stetig differenzierbare Kurve zwischen $\boldsymbol{P}_1$ und $\boldsymbol{P}_2$ und $L(f)$ die Bogenlänge von f ist. Der Abstand $\mathrm{g}(\boldsymbol{P}_1, \boldsymbol{P}_2)$ entspricht

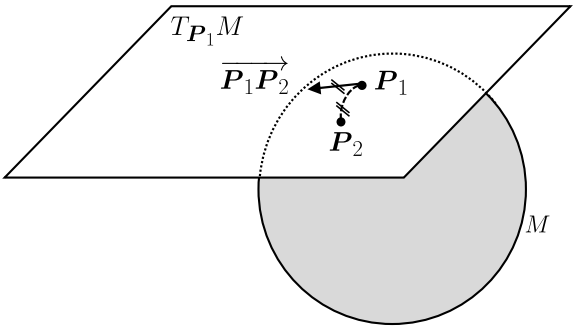


Abbildung 4.1: Die 2-Sphäre $\boldsymbol{S}^2$ ist ein Beispiel für eine 2-dimensionale differenzierbare Mannigfaltigkeit. Sie ist lokal — in einer kleinen Umgebung um jeden Punkt $\boldsymbol{P} \in M$ der Mannigfaltigkeit — diffeomorph zu einer Tangentialebene $T_{\boldsymbol{P}}M$, $\boldsymbol{P} \in M$, in der geometrische Berechnungen durchgeführt werden können, da die Mannigfaltigkeit mit einer riemannschen Metrik versehen werden kann.

dann der Länge einer global kürzesten Verbindungskurve, der sich aus folgender Gleichung ergibt:

$$\mathrm{g}(\boldsymbol{P}_1, \boldsymbol{P}_2) = \inf \{L(f) \,|\, f : [0,1] \to M, f(0) = \boldsymbol{P}_1, f(1) = \boldsymbol{P}_2\} \,. \quad (4.3)$$

Abbildung 4.1 veranschaulicht eine (riemannsche) Mannigfaltigkeit M. In dem dargestellten Beispiel soll die Kugeloberfläche betrachtet werden. Sei die Oberfläche die 2-Sphäre $\boldsymbol{S}^2$, dann entspricht sie einer 2-dimensionalen Mannigfaltigkeit, die in den euklidischen Raum $\mathbb{R}^3$ eingebettet[1] ist. Besitzt diese Mannigfaltigkeit zudem eine riemannsche Metrik, spricht man von einer riemannschen Mannigfaltigkeit. Die riemannsche Metrik ermöglicht es, Winkel und Entfernungen zwischen Punkten der Mannigfaltigkeit zu bestimmen. Der geodätische Abstand $\mathrm{g}(\boldsymbol{P}_1, \boldsymbol{P}_2)$ von zwei Punkten $\boldsymbol{P}_1$ und $\boldsymbol{P}_2$ auf der Sphäre entspricht der Länge einer kürzesten Verbindungskurve (minimierenden Geodätische) zwischen den beiden Punkten. Liegen die

[1] Das Thema Einbettung ist u.a. für das Auffinden niedrigdimensionaler Mannigfaltigkeiten relevant, die in Räumen höherer Dimensionen eingebettet sind (vgl. Abschnitt 4.2). Im Rahmen dieser Arbeit stellt sich diese Problematik bei der Personendetektion und -wiedererkennung.

Punkte $\boldsymbol{P}_1$ und $\boldsymbol{P}_2$ ausreichend nahe beieinander, dann entspricht der geodätische Abstand $\mathrm{g}(\boldsymbol{P}_1, \boldsymbol{P}_2)$ der euklidischen Norm von $||\overrightarrow{\boldsymbol{P}_1\boldsymbol{P}_2}||_2$ im Tangentialraum $T_{\boldsymbol{P}_1}M$ am Punkt $\boldsymbol{P}_1$:

$$\lim_{\boldsymbol{P}_2 \to \boldsymbol{P}_1} \left(\mathrm{g}(\boldsymbol{P}_1, \boldsymbol{P}_2) - ||\overrightarrow{\boldsymbol{P}_1\boldsymbol{P}_2}||_2 \right) = 0\,. \tag{4.4}$$

Es muss dabei allerdings beachtet werden, dass in dem gegebenen Beispiel der 2-Sphäre der Diffeomorphismus[2], der die Punkte aus den Tangentialräumen in die Mannigfaltigkeit bzw. umgekehrt abbildet, keine Isometrie ist. Das gilt im Allgemeinen für die meisten Mannigfaltigkeiten. Das bedeutet, dass der geodätische Abstand zwischen zwei weit auseinanderliegenden Punkten aus der Summe der Längen von Geradenstücken zwischen ausreichend nahe beieinanderliegenden Zwischenpunkten bestimmt werden muss, der sich aus der Abstandsfunktion (4.3) ergibt. Für weiterführende Grundlagen wird an dieser Stelle auf [Ber18] verwiesen.

4.1.1 Affin-invariante riemannsche Metrik

Gemäß dem Beispiel der 2-Sphäre $\boldsymbol{S}^2$ in Abbildung 4.1, entspricht der geodätische Abstand zweier positiv definiter Kovarianzdeskriptoren des Sym_n^+ der Länge einer kürzesten Verbindungskurve zwischen den beiden Deskriptoren. Der Raum der Kovarianzdeskriptoren Sym_n^+ ist eine riemannsche Mannigfaltigkeit [Pen06a]. Eine geeignete affin-invariante riemannsche Metrik zur Bestimmung des Abstands wird in [Pen06b] hergeleitet, die unabhängig davon auch in [För99, Fle04, Len04, Moa05, Len06] vorgeschlagen wird. Sie hat sich in den letzten Jahren als Metrik für den Sym_n^+ etabliert (siehe z.B. [Bat05, Tuz06, Fle07]). Es ist die erste bekannte Metrik für den Raum der Kovarianzdeskriptoren Sym_n^+, die unter affinen Transformationen und Inversionen invariant bleibt [För99]. Im Rahmen dieser Arbeit wird ebenfalls diese Metrik zur Bestimmung der geodätischen Abstände zwischen positiv definiten Kovarianzdeskriptoren verwendet.

Für jedes $\boldsymbol{\Sigma} \in \mathrm{Sym}_n^+$ und $\boldsymbol{a}, \boldsymbol{b} \in T_{\boldsymbol{\Sigma}}\,\mathrm{Sym}_n^+$ ist durch die Formel

$$\langle \boldsymbol{a} \,|\, \boldsymbol{b} \rangle_{\boldsymbol{\Sigma}} := \mathrm{Spur}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\, \boldsymbol{a}\, \boldsymbol{\Sigma}^{-1}\, \boldsymbol{b}\, \boldsymbol{\Sigma}^{-\frac{1}{2}} \right) \tag{4.5}$$

[2]Ein Diffeomorphismus ist eine bijektive, stetig differenzierbare Abbildung, deren Umkehrabbildung ebenfalls stetig differenzierbar ist.

ein Skalarprodukt gegeben [Pen06a]. Die Skalarprodukte hängen differenzierbar von $\boldsymbol{\Sigma}$ ab und definieren daher eine riemannsche Metrik [Pen06a], die im Folgenden als $\mathrm{g}(\cdot)$ bezeichnet wird.

Eine riemannsche Mannigfaltigkeit ist lokal diffeomorph. Damit ist der Sym_n^+ in jedem Punkt $\boldsymbol{\Sigma} \in \mathrm{Sym}_n^+$ lokal diffeomorph zu einem $\frac{n(n+1)}{2}$-dimensionalen Tangentialraum $T_{\boldsymbol{\Sigma}}\,\mathrm{Sym}_n^+$. Eine Abbildung, die jeden Punkt $\boldsymbol{a} \in T_{\boldsymbol{\Sigma}}\,\mathrm{Sym}_n^+$ in den Sym_n^+ abbildet, ist die riemannsche Exponentialabbildung [Pen06a]. Für die riemannsche Metrik $\mathrm{g}(\cdot)$ ist die riemannsche Exponentialabbildung gegeben durch

$$\begin{aligned} \exp_{\boldsymbol{\Sigma}} : T_{\boldsymbol{\Sigma}}\,\mathrm{Sym}_n^+ &\rightarrow \mathrm{Sym}_n^+ \qquad \text{mit} \\ \boldsymbol{a} \mapsto \exp_{\boldsymbol{\Sigma}}(\boldsymbol{a}) &= \boldsymbol{\Sigma}^{\frac{1}{2}}\, \exp\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\, \boldsymbol{a}\, \boldsymbol{\Sigma}^{-\frac{1}{2}}\right) \boldsymbol{\Sigma}^{\frac{1}{2}}\,, \end{aligned} \tag{4.6}$$

wobei exp die gewöhnliche Exponentialabbildung von Matrizen ist [Pen06a].

Dieser Diffeomorphismus besitzt die vorteilhafte Eigenschaft, dass er entlang der Geodätischen durch $\boldsymbol{\Sigma}$ eine Isometrie ist und damit die Punkte $\boldsymbol{\Sigma} \in \mathrm{Sym}_n^+$, die auf einer Geodäten liegen, längenerhaltend abbildet. Außerdem hat er die praktische Eigenschaft, dass er im Sym_n^+ global ist, d.h. $\exp_{\boldsymbol{\Sigma}_1}$ ist in jedem Punkt $\boldsymbol{\Sigma} \in \mathrm{Sym}_n^+$ ein globaler Diffeomorphismus und es existiert eine Umkehrabbildung, die global — in jedem Punkt $\boldsymbol{\Sigma} \in \mathrm{Sym}_n^+$ — eindeutig definiert ist [Pen06a]. Die Umkehrabbildung ist mit dem gewöhnlichen Logarithmusoperator log von Matrizen definiert durch

$$\begin{aligned} \log_{\boldsymbol{\Sigma}_0} : \mathrm{Sym}_n^+ &\rightarrow T_{\boldsymbol{\Sigma}_0}\,\mathrm{Sym}_n^+ \qquad \text{mit} \\ \boldsymbol{\Sigma} \mapsto \log_{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma}) &= \boldsymbol{\Sigma}_0^{\frac{1}{2}}\, \log\left(\boldsymbol{\Sigma}_0^{-\frac{1}{2}}\, \boldsymbol{\Sigma}\, \boldsymbol{\Sigma}_0^{-\frac{1}{2}}\right) \boldsymbol{\Sigma}_0^{\frac{1}{2}}\,, \end{aligned} \tag{4.7}$$

wobei $\boldsymbol{\Sigma}_0$ der Basispunkt am Tangentialraum $T_{\boldsymbol{\Sigma}_0}\,\mathrm{Sym}_n^+$ ist.

Aufgrund der genannten Eigenschaften des Diffeomorphismus $\exp_{\boldsymbol{\Sigma}}$ kann der geodätische Abstand zwischen zwei beliebigen positiv definiten Kovarianzmatrizen $\boldsymbol{\Sigma}_0$ und $\boldsymbol{\Sigma}$ einfach berechnet werden. Der geodätische Abstand $\mathrm{g}(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma})$ entspricht dank der Eigenschaften der euklidischen Länge des zu $\boldsymbol{\Sigma}$ gehörenden Tangentialvektors in $T_{\boldsymbol{\Sigma}_0}\,\mathrm{Sym}_n^+$:

$$\mathrm{g}(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma})^2 = ||\log_{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})||^2_{\boldsymbol{\Sigma}_0}\,. \tag{4.8}$$

Durch Einsetzen der Gleichung (4.7) in Gleichung (4.5) ergibt sich folgende Gleichung für die Berechnung des geodätischen Abstands aus Gleichung (4.8):

$$\begin{aligned} || \log_{\mathbf{\Sigma}_0}(\mathbf{\Sigma})||^2_{\mathbf{\Sigma}_0} &= \langle \log_{\mathbf{\Sigma}_0}(\mathbf{\Sigma}) \mid \log_{\mathbf{\Sigma}_0}(\mathbf{\Sigma}) \rangle_{\mathbf{\Sigma}_0} \\ &= \text{Spur}\left(\mathbf{\Sigma}_0^{-\frac{1}{2}} \ \log_{\mathbf{\Sigma}_0}(\mathbf{\Sigma}) \ \mathbf{\Sigma}_0^{-1} \ \log_{\mathbf{\Sigma}_0}(\mathbf{\Sigma}) \ \mathbf{\Sigma}_0^{-\frac{1}{2}} \right) \\ &= \text{Spur}\left(\log\left(\mathbf{\Sigma}_0^{-\frac{1}{2}} \, \mathbf{\Sigma} \, \mathbf{\Sigma}_0^{-\frac{1}{2}} \right) \ \log\left(\mathbf{\Sigma}_0^{-\frac{1}{2}} \, \mathbf{\Sigma} \, \mathbf{\Sigma}_0^{-\frac{1}{2}} \right) \right) \\ &= \text{Spur}\left(\log^2\left(\mathbf{\Sigma}_0^{-\frac{1}{2}} \, \mathbf{\Sigma} \, \mathbf{\Sigma}_0^{-\frac{1}{2}} \right) \right) . \end{aligned} \tag{4.9}$$

4.1.2 Mittelwertberechnung

Die Mittelwertberechnung kann verwendet werden, um beispielsweise mehrere nahe beieinander liegende Kovarianzdeskriptoren zu einem Deskriptor zusammenzufassen. In Kapitel 6 werden für das Tracking einzelner Personen aus mehreren Deskriptoren ein Mittelwert für die Personenrepräsentation bestimmt, um u.a. Änderungen der Personenerscheinung und Beleuchtungsänderungen zu berücksichtigen. Bei der erscheinungsbasierten Personenwiedererkennung in Kapitel 7 werden während der Berechnung von Personenrepräsentationen aus Tracklets ebenfalls mehrere *ähnliche* Kovarianzdeskriptoren zusammengefasst, um Teile von Personen-Tracklets zu repräsentieren. Aber auch in Fällen, in denen Kovarianzdeskriptoren *weit* auseinander liegen, kann eine mittelwertbasierte Repräsentation von Kovarianzdeskriptoren vorteilhaft sein, wie der lernbasierte Ansatz für die Bestimmung von Kovarianzdeskriptoren zur Repräsentation von Personen aus mehreren Einzelbildern in Abschnitt 8.1 bekräftigt.

Es gibt mehrere Definitionen eines empirischen Mittelwertes für positiv definite symmetrische Matrizen [Pen06b]. Im Rahmen dieser Arbeit wird der *Karcher*- bzw. *Fréchet*-Mittelwert verwendet, der die Summe der quadratischen geodätischen Abstände zwischen dem Mittelwert und den Kovarianzdeskriptoren minimiert [Kar77, Pen06c]. Nach [Sko84] hat der Sym_n^+ eine nicht-positive Krümmung. Bei solchen hyperbolischen Mannigfaltigkeiten lässt sich der Mittelwert bestimmen. Im Sym_n^+ ist nach [Ken90] der *Karcher*-Mittelwert aufgrund der Eigenschaft von [Sko84] eindeutig.

Der Mittelwert kann iterativ mit dem Gauß-Newton-Verfahren bestimmt werden [Pen06c], das in einem Tangentialraum $T_{\bar{\mathbf{\Sigma}}} \operatorname{Sym}_n^+$ durchgeführt wird. Dazu werden in einem Iterationsschritt die Kovarianzdeskriptoren $\mathcal{A} = \{\mathbf{\Sigma}_i\}$, $i = 1, \ldots, n_a$, aus denen ein Mittelwert $\bar{\mathbf{\Sigma}} \in \operatorname{Sym}_n^+$ bestimmt werden soll, mit der Umkehrabbildung der riemannschen Exponentialabbildung (Gleichung (4.7)) in den Tangentialraum $T_{\bar{\mathbf{\Sigma}}^t} \operatorname{Sym}_n^+$ des Schätzwerts der aktuellen Iteration abgebildet. Aus den resultierenden Tangentialvektoren wird dann ein euklidischer Mittelwert berechnet und mittels der riemannschen Exponentialabbildung (Gleichung (4.6)) zurück in den Sym_n^+ abgebildet. Damit ergibt sich folgende Gleichung für den Iterationsschritt $t + 1$:

$$\bar{\mathbf{\Sigma}}^{t+1} := \exp_{\bar{\mathbf{\Sigma}}^t} \left(\frac{1}{n_a} \sum_{i=1}^{n_a} \log_{\bar{\mathbf{\Sigma}}^t}(\mathbf{\Sigma}_i) \right) . \tag{4.10}$$

Für den ersten Iterationsschritt ($t = 0$) kann ein beliebiger Kovarianzdeskriptor aus der Menge $\mathcal{A}$ eingesetzt werden [Pen06a].

Der in Gleichung (4.10) angegebene Iterationsschritt wird solange wiederholt, bis die Bedingung $\mathrm{g}(\bar{\mathbf{\Sigma}}^{t+1}, \bar{\mathbf{\Sigma}}^t) < \psi$ erfüllt ist, wobei ψ nahe 0 ist. In der Regel ist diese Bedingung nach weniger als fünf Iterationsschritten erfüllt. $\bar{\mathbf{\Sigma}}^{t+1}$ entspricht dann dem Mittelwert $\bar{\mathbf{\Sigma}}$.

4.1.3 Empirische Kovarianzmatrix

Viele gewöhnliche statistische Methoden können durch Verwendung der riemannschen Exponentialabbildung (Gleichung (4.7)) an einem Mittelwert $\bar{\mathbf{\Sigma}} \in \operatorname{Sym}_n^+$ von Kovarianzdeskriptoren für statistische Betrachtungen im Sym_n^+ verallgemeinert werden [Pen06c]. Beispielsweise lässt sich eine Wahrscheinlichkeitsdichtefunktion finden, die die Information bei einem eindeutigen Mittelwert $\bar{\mathbf{\Sigma}}$ und empirischer Kovarianzmatrix $\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}}$ minimiert. Die generalisierte Normalverteilung im Sym_n^+ ist gemäß [Pen06c] über die Normalverteilungen in den Tangentialräumen des Sym_n^+ definiert.

Die generalisierte Normalverteilung $\mathcal{N}(\bar{\mathbf{\Sigma}}, \boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})$ ist durch die Dichtefunktion

$$p_{\bar{\mathbf{\Sigma}}}(\mathbf{\Sigma}) = k \cdot \exp\left(-\frac{1}{2} \log_{\bar{\mathbf{\Sigma}}}(\mathbf{\Sigma}) \quad \boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}}^{-1} \quad \log_{\bar{\mathbf{\Sigma}}}(\mathbf{\Sigma}) \right) \tag{4.11}$$

gegeben, wobei k die Normalisierungskonstante ist. Für weiterführende Details wird an dieser Stelle auf die Arbeiten [Pen99, Pen06c] verwiesen.

Die empirische Kovarianzmatrix $\boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}}$ für die Menge der Kovarianzdeskriptoren $\mathcal{B} = \{\boldsymbol{\Sigma}_i\}$, $i = 1, \ldots, n_b$, $\boldsymbol{\Sigma}_i \in \mathrm{Sym}_n^+$, mit dem Mittelwert $\bar{\boldsymbol{\Sigma}}$ berechnet sich durch

$$\boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}} := \frac{1}{n_b - 1} \sum_{i=1}^{n_b} \mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}} \left(\overrightarrow{\bar{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_i} \right) \mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}} \left(\overrightarrow{\bar{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_i} \right)^T , \tag{4.12}$$

wobei $\overrightarrow{\bar{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_i}$ ein Tangentialvektor des Tangentialraums $T_{\bar{\boldsymbol{\Sigma}}} \mathrm{Sym}_n^+$ und $\mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}}$ ein Isomorphismus zwischen dem Tangentialraum $T_{\bar{\boldsymbol{\Sigma}}} \mathrm{Sym}_n^+$ und $\mathbb{R}^{\frac{n(n+1)}{2}}$ ist (vgl. [Pen06a]):

$$\mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}}(\overrightarrow{\bar{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}_i}) = \mathrm{Vec}_{\boldsymbol{E}} \left(\log \left(\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_i \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \right) \right) . \tag{4.13}$$

$\boldsymbol{E}$ ist die Einheitsmatrix und $\mathrm{Vec}_{\boldsymbol{E}}(\boldsymbol{S})$ ein Operator, der die Elemente $\{s_{i,j}\}_{i,j=1,\ldots,n}$ einer $n \times n$ Matrix $\boldsymbol{S}$ in einen Vektor an die Stellen $(i \cdot n + j)$ transformiert (vgl. [Pen06c]):

$$\begin{aligned} \mathrm{Vec}_{\boldsymbol{E}}(\boldsymbol{S}) := (s_{1,1}, \sqrt{2}s_{1,2}, s_{2,2}, \sqrt{2}s_{1,3}, \sqrt{2}s_{2,3}, s_{3,3}, \ldots \\ \ldots, \sqrt{2}s_{1,n}, \ldots, \sqrt{2}s_{(n-1),n}, s_{n,n})^T . \end{aligned} \tag{4.14}$$

Die empirische Kovarianzmatrix ist im Rahmen dieser Arbeit für die Berechnung der *Mahalanobis*-Distanz erforderlich. Die Definition der Mahalanobis-Distanz für den Sym_n^+ ist im nächsten Abschnitt gegeben.

4.1.4 Mahalanobis-Distanz

Die Mahalanobis-Distanz wird für das Tracking von Einzelpersonen zur Bestimmung des Abstands zwischen einer Beobachtung im Sym_n^+ (Kovarianzdeskriptor) und einer Normalverteilung von Kovarianzdeskriptoren verwendet (siehe Abschnitt 6.2).

Die im $\mathbb{R}^n$ definierte Mahalanobis-Distanz kann für den Sym_n^+ verallgemeinert werden. Nach [Pen06c] ist sie über den empirischen Mittelwert

$\bar{\mathbf{\Sigma}} \in \mathrm{Sym}_n^+$ und der empirischen Kovarianzmatrix $\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}}$ wie folgt definiert:

$$\Omega_{(\bar{\mathbf{\Sigma}},\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})}(\mathbf{\Sigma}) := \sqrt{\mathrm{Vec}_{\bar{\mathbf{\Sigma}}}(\overrightarrow{\bar{\mathbf{\Sigma}}\mathbf{\Sigma}})^T \quad \boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}}^{-1} \quad \mathrm{Vec}_{\bar{\mathbf{\Sigma}}}(\overrightarrow{\bar{\mathbf{\Sigma}}\mathbf{\Sigma}})}\,. \tag{4.15}$$

Der Operator $\mathrm{Vec}_{\bar{\mathbf{\Sigma}}}$ entspricht dem durch die Gleichung (4.13) gegebenen Isomorphismus. Die Mahalanobis-Distanz $\Omega_{(\bar{\mathbf{\Sigma}},\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})}$ ist für beliebige Verteilungen einer Zufallsvariable wohldefiniert [Pen06c]. Der empirische Mittelwert $\bar{\mathbf{\Sigma}}$ und die zugehörige empirische Kovarianzmatrix $\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}}$ lassen sich gemäß den beiden vorherigen Abschnitten berechnen.

χ^2-Test

Die Überprüfung, ob ein Kovarianzdeskriptor aus einer bestimmten Normalverteilung stammt, erfolgt mittels χ^2-Test im Sym_n^+, mit dem Ausreißer detektiert werden können. Im Rahmen dieser Arbeit wird ein ähnlicher Ansatz für die Aktualisierungsstrategie beim Kovarianz-Trackingverfahren verfolgt.

Sei $\mathbf{\Sigma} \sim \mathcal{N}(\bar{\mathbf{\Sigma}}, \boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})$ eine normalverteilte Zufallsmatrix. $\Omega^2_{(\bar{\mathbf{\Sigma}},\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})}(\mathbf{\Sigma})$ sollte dann χ^2_k-verteilt sein, mit $k = \frac{n(n+1)}{2}$, wenn die Beobachtung korrekt ist [Pen06c]. Damit lässt sich testen, ob eine Beobachtung zu einer Normalverteilung gehört oder als Ausreißer betrachtet werden muss. Dazu wird zunächst die Wahrscheinlichkeit berechnet, dass

$$\chi^2_k = \Omega^2_{(\bar{\mathbf{\Sigma}},\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})}(\mathbf{\Sigma}) = \mathrm{Vec}_{\bar{\mathbf{\Sigma}}}(\overrightarrow{\bar{\mathbf{\Sigma}}\mathbf{\Sigma}})^T \quad \boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}}^{-1} \quad \mathrm{Vec}_{\bar{\mathbf{\Sigma}}}(\overrightarrow{\bar{\mathbf{\Sigma}}\mathbf{\Sigma}}) \leq \alpha^2\,. \tag{4.16}$$

Gehört die Beobachtung $\mathbf{\Sigma}$ zur Normalverteilung, ist $\Omega_{(\bar{\mathbf{\Sigma}},\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})}$ mit einer Wahrscheinlichkeit $\gamma = Pr\{\chi^2_k \leq \alpha^2\}$ kleiner als α^2. Zu dieser Überprüfung muss ein Konfidenzniveau γ gewählt werden (üblicherweise $\geq 95\%$) und ein Wert $\alpha(\gamma)$ gesucht werden, so dass die Gleichung $\gamma = Pr\{\chi^2_k \leq \alpha^2\}$ gilt. Ist $\Omega^2_{(\bar{\mathbf{\Sigma}},\boldsymbol{Cov}_{\bar{\mathbf{\Sigma}}})}(\mathbf{\Sigma}) \leq \alpha^2$, wird davon ausgegangen, dass die Beobachtung zur Normalverteilung gehört, andernfalls wird die Beobachtung als Ausreißer behandelt.

4.1.5 Praktische Aspekte

Dieser Abschnitt schließt das Unterkapitel über den Sym_n^+ mit zwei wichtigen praktischen Aspekten: Die Sicherstellung der positiven Definitheit der Kovarianzdeskriptoren und einer alternativen Metrik zum Vergleich der Deskriptoren, die beispielsweise aus Laufzeitgründen in Betracht gezogen werden kann.

Positive Definitheit

Der Sym_n^+ ist für reelle positiv definite Kovarianzdeskriptoren definiert und berücksichtigt keine semidefiniten Kovarianzdeskriptoren. Deswegen werden in der Praxis die Kovarianzdeskriptoren normalerweise auf positive Definitheit überprüft und im Falle von Semidefinitheit wird die Kovarianzmatrix entweder nicht berücksichtigt oder mit einem *Semidefiniten-Programmierung-Löser* die nächste positiv definite Kovarianzmatrix zu dieser bestimmt. Die Sicherstellung der positiven Definitheit kann beispielsweise mittels der *Cholesky-Zerlegung* erfolgen, indem überprüft wird, ob die Zerlegung durchführbar ist (sie ist nur für symmetrische positiv definite Matrizen berechenbar) oder einfach durch Überprüfung der Varianzen, die alle durchweg positiv sein müssen. Im Rahmen dieser Arbeit wird ein kleiner konstanter Wert auf die Varianzen addiert, um semidefinite Kovarianzdeskriptoren auszuschließen.

log-euklidische Metrik

Die in Abschnitt 4.1.1 vorgestellte affin-invariante riemannsche Metrik bildet zusammen mit dem hier vorgestellten Rahmenwerk ein sehr mächtiges und nützliches Werkzeug zur Verarbeitung und Anwendung von Kovarianzdeskriptoren [Ars06]. Allerdings sind die Algorithmen, die auf dieser Metrik basieren, sehr komplex und langsam, wie z.B. in der Veröffentlichung von Arsigny et al. ersichtlich ist. Um eine einfachere und schnellere Verarbeitung von Kovarianzdeskriptoren zu ermöglichen, schlagen die Autoren eine log-euklidische Metrik vor, mit der die Abstandsberechnung durchschnittlich in einem Siebtel der Zeit durchgeführt werden kann. Die Idee dabei ist, zunächst die Matrixlogarithmen der Kovarianzdeskriptoren

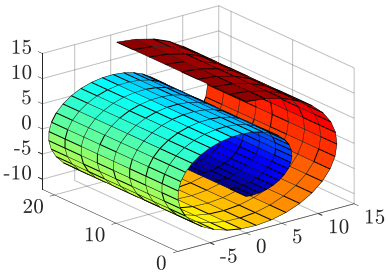
zu bestimmen, mit denen euklidisch weitergerechnet werden kann. Die riemannschen Operationen können dann durch euklidische Operationen ersetzt werden. Allerdings ist bei der log-euklidischen Metrik z.B. die affine Invarianz nicht gegeben, so dass hier zusätzliche Mechanismen für diese Anforderung eingesetzt werden müssen. In der Praxis erzielen beide Metriken ähnliche Ergebnisse, weswegen aus Laufzeitgründen es durchaus Sinn macht, die affin-invariante Metrik durch die log-euklidische Metrik zu ersetzen. Es sollte dabei nur sichergestellt werden, dass keine Leistungseinbußen in puncto Genauigkeit entstehen. Im Rahmen dieser Arbeit liegt aufgrund ihrer guten theoretischen Eigenschaften, der Fokus auf der affin-invarianten Metrik. Für weitere Details, was die log-euklidische Metrik betrifft, wird deshalb an dieser Stelle auf [Ars06] verwiesen.

4.2 Nichtlineare Dimensionsreduktion

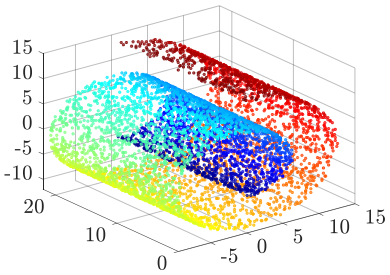
Das Ziel der Dimensionsreduktion ist das Auffinden niedrigdimensionaler Repräsentationen von hochdimensionalen Daten bei minimalem Informationsverlust. Dimensionsreduktionsalgorithmen bilden dazu D-dimensionale Eingangsdaten vom $\mathbb{R}^D$ in einen Raum $\mathbb{R}^d$ wesentlicher niedriger Dimension ab ($d \ll D$). Wesentliches Entwurfskriterium der Algorithmen ist dabei, dass die (lokalen) Abstände zwischen den Punkten bestmöglich erhalten bleiben. Anwendungen der Dimensionsreduktion sind beispielsweise die Visualisierung hochdimensionaler Daten, Datenkompression und Datenvorverarbeitung für Klassifikationsverfahren. Im Rahmen dieser Arbeit wird die Dimensionsreduktion zur Merkmalsextraktion eingesetzt, die bei der Personendetektion und -wiedererkennung zur Anwendung kommt.

Der wohl bekannteste Dimensionsreduktionsalgorithmus ist die Hauptkomponentenanalyse [Pea01, Lee07]. Bei dieser Methode werden die Eingabedaten so in einen niedrigdimensionalen Raum projiziert, dass der Informationsverlust möglichst gering ist. Die Hauptkomponentenanalyse ist eine lineare Methode und eignet sich für Daten, die auf einer linearen Mannigfaltigkeit liegen.

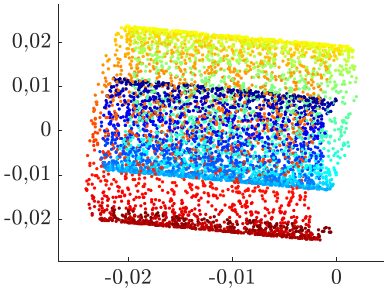
Für Daten, die auf einer nichtlinearen Mannigfaltigkeit liegen, ist der Ansatz linearer Methoden ungeeignet, wie das Beispiel in Abbildung 4.2 veranschaulicht. Die Fläche, die in Abbildung 4.2a zu sehen ist, ist unter der Bezeichnung *Swiss Roll* bekannt. Es handelt sich dabei um eine nichtlineare zweidimensionale Mannigfaltigkeit, die in den $\mathbb{R}^3$ eingebettet ist. Diese Mannigfaltigkeit motiviert sehr gut den Einsatz nichtlinearer Dimensionsreduktionsmethoden. Es ist einfach zu erkennen, dass es keine lineare Dimensionsreduktion gibt, welche die geodätischen Abstände erhält. Das Ziel der Dimensionsreduktion für dieses Beispiel ist das Auffinden einer 2-dimensionalen Repräsentationen des Swiss Roll Datensatzes, so dass die lokalen Abstände erhalten bleiben, was dem Aufrollen der Swiss Roll entspricht und nur durch nichtlineare Dimensionsreduktionsmethoden erzielt werden kann. Abbildung 4.2c zeigt das Ergebnis einer linearen Reduktion der Dimension auf 2 mittels Hauptkomponentenanalyse, wobei die Verhältnisse der geodätischen Abstände nicht erhalten werden konnten (im abgebildeten 2-dimensionalen Raum können die geodätischen Abstände durch die euklidischen Abstände approximiert werden). Der nichtlinea-



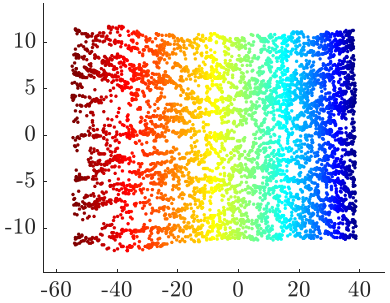
(a) Swiss Roll



(b) Swiss Roll Datensatz



(c) Hauptkomponentenanalyse



(d) Isomap

Abbildung 4.2: Die Swiss Roll ist ein Beispiel für eine nichtlineare zweidimensionale Mannigfaltigkeit, die in den $\mathbb{R}^3$ eingebettet ist und für die es keine lineare Einbettung in den $\mathbb{R}^2$ gibt. Swiss Roll Mannigfaltigkeit **(a)**, Datensatz mit 6.000 Punkten, die von der Swiss Roll Fläche abgetastet wurden **(b)**, 2-dimensionale Einbettung des Swiss Roll Datensatzes mittels Hauptkomponentenanalyse **(c)**, 2-dimensionale Einbettung des Swiss Roll Datensatzes mittels Isomap **(d)**.

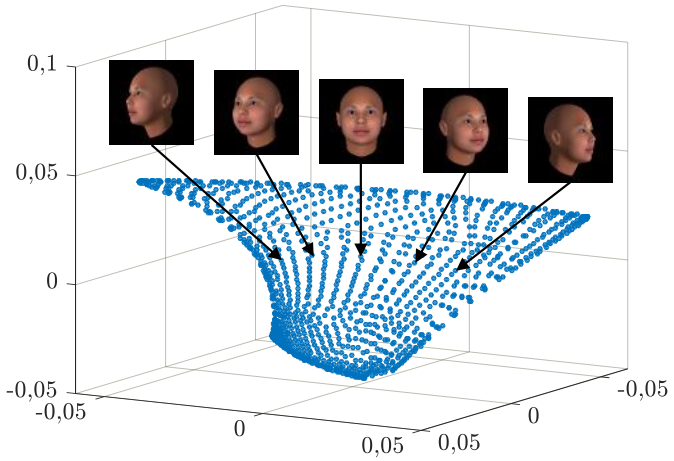


Abbildung 4.3: Mittels LE (Laplacian Eigenmaps) in den $\mathbb{R}^3$ eingebettete 3-dimensionale Mannigfaltigkeit einer synthetisch erzeugten Bildsequenz von Gesichtern eines Individuums.

re Dimensionsreduktionsalgorithmus Isomap [Ten00] konnte hingegen die Abstände zwischen den Punkten sehr gut erhalten.

In der Bildauswertung, insbesondere bei der Auswertung von Videosequenzen von Personen, liegen die Bilddaten in vielen Fällen auf nichtlinearen Mannigfaltigkeiten, für die lineare Ansätze im Allgemeinen — wie bei dem vorangegangen Beispiel — keine geeignete niedrigdimensionale Repräsentation finden. Ein Beispiel dafür ist der in Abbildung 4.3 dargestellte Auszug aus einer Bildsequenz von synthetisch erzeugten Gesichtern eines Individuums, die aus verschiedenen Perspektiven zu sehen sind. Für die unterschiedlichen Ansichten wurde ein Gesicht um seine Vertikal-, Längs- und Querachse rotiert, wobei jedes Mal eine Frontalbeleuchtung gewählt wurde, so dass die Bilder eine Mannigfaltigkeit mit einer intrinsischen Dimension von 3 erzeugen (die intrinsische Dimension entspricht in dem gegebenen Beispiel der Anzahl der Freiheitsgrade des Gesichts in der Sequenz). Obwohl sich die Mimik in der gesamten Bildsequenz nicht ändert, liegen die Gesichtsbilder auf einer gekrümmten Mannigfaltigkeit. Bildet

man die Grauwerte beispielsweise mit dem in Gleichung (4.14) definierten Vec-Operator in einen Vektor ab, lässt sich erkennen, dass die Punkte auf — bzw. aufgrund von messtechnischen Fehlern bei Gesichtsaufnahmen nahe bei — einer nichtlinearen Mannigfaltigkeit liegen. Das Diagramm in der Abbildung 4.3 visualisiert dies anhand der Einbettung der durch die Gesichtsbilder erzeugten Mannigfaltigkeit in den $\mathbb{R}^3$ mittels dem Dimensionsreduktionsalgorithmus LE [Bel03] (das Verfahren LE wird ausführlich weiter unten in diesem Abschnitt vorgestellt).

Anhand synthetisch erzeugter Gesichtsbilder lassen sich die vorhandenen Nichtlinearitäten sehr gut veranschaulichen. Aber auch Bildausschnitte, die nicht synthetisch erstellt sind und Personen vollständig abbilden, liegen sehr oft auf einer Mannigfaltigkeit, wie Abschnitt 4.2.2 exemplarisch aufzeigt.

Eine Gruppe von Methoden zur nichtlinearen Dimensionsreduktion sind Manifold Learning (MaL) Algorithmen, die für Daten geeignet sind, die auf einer nichtlinearen Mannigfaltigkeit liegen. Die Algorithmen zeichnen sich insbesondere dadurch aus, dass sie beim Auffinden niedrigdimensionaler Repräsentationen von Mannigfaltigkeiten, die in sehr hochdimensionalen Räumen eingebettet sind, sehr gut die geodätischen Abstände zwischen den Datenpunkten erhalten können. Im nächsten Abschnitt wird das Thema MaL näher behandelt, mit Schwerpunkt auf dem in dieser Arbeit verwendeten Algorithmus.

4.2.1 Manifold Learning

Im Rahmen dieser Arbeit wird ein MaL Algorithmus zur Dimensionsreduktion und Merkmalsextraktion verwendet. MaL Methoden basieren auf der Beobachtung, dass niedrigdimensionale Mannigfaltigkeiten meistens in hochdimensionale Räume eingebettet sind.

Das MaL Problem lässt sich wie folgt formal beschreiben. Sei $\mathcal{X} = \{\boldsymbol{x}_i\}, i = 1, \ldots, n$, $\boldsymbol{x}_i \in \mathbb{R}^D$ eine Menge mit n Punkten, für die eine niedrigdimensionale Repräsentation gefunden werden soll. Unter der Annahme, dass die Punkte der Menge $\mathcal{X}$ auf einer d-dimensionalen Mannigfaltigkeit liegen, die in den $\mathbb{R}^D$ eingebettet ist ($D \gg d$) und die Mannigfaltigkeit durch

einen Homöomorphismus[3] h beschrieben werden kann (bzw. Diffeomorphismus[4] bei einer differenzierbaren Mannigfaltigkeit), ist das Ziel von MaL eine niedrigdimensionale Repräsentation $\mathcal{Y} = \{\boldsymbol{y}_i\}$, $i = 1, \ldots, n$, $\boldsymbol{y}_i \in \mathbb{R}^d$ zu finden, so dass $\boldsymbol{y}_i = h(\boldsymbol{x}_i)$ für $i = 1, \ldots, n$ gilt. Das MaL Problem lässt sich auch lösen, wenn sich die Mannigfaltigkeit durch mehrere lokale Homöomorphismen bzw. Diffeomorphismen beschreiben lässt.

Die Dimension d muss dabei in der Regel vorab vorgegeben werden, was ein generelles Problem bei der Dimensionsreduktion darstellt. In den meisten Fällen ist die intrinsische Dimension der Mannigfaltigkeit, welche der Anzahl der unabhängigen Variablen einer Datenmenge entspricht, unbekannt, so dass diese geschätzt werden muss. Die Schätzung der intrinsischen Dimension ist ein essentieller Schritt in der Dimensionsreduktion, da für eine möglichst genaue niedrigdimensionale Repräsentation einer Datenmenge die vorgegebene Dimension d im Allgemeinen möglichst der tatsächlichen intrinsischen Dimension der Mannigfaltigkeit, die der Datenmenge zugrunde liegt, entsprechen sollte. Verschiedene Möglichkeiten zur Schätzung der intrinsischen Dimension werden in [Lee07] behandelt.

MaL Algorithmen werden seit Anfang der 2000er Jahre, nach Veröffentlichung der beiden MaL Algorithmen *Isometric Feature Mapping* (Isomap) [Ten00] und *Locally Linear Embedding* (LLE) [Row00], gerne zur nichtlinearen Dimensionsreduktion eingesetzt, da sie bei der Dimensionsreduktion sehr gut die geodätischen Abstände der Datenpunkte erhalten. Die Algorithmen können in zwei Gruppen eingeteilt werden: lokale und globale Verfahren. Globale Ansätze wie beispielsweise Isomap versuchen die globalen geodätischen Abstände zu erhalten, während lokale Verfahren wie LLE [Row00] oder LE sich darauf fokussieren, die lokalen Abstände zwischen Nachbarn zu erhalten. Der wesentliche Vorteil der globalen Verfahren ist, dass sie dazu tendieren, eine originalgetreuere Repräsentation der Eingabedaten zu lernen. Ein Nachteil dieser Ansätze ist allerdings, dass sie bei stark gekrümmten und geodätisch nicht-konvexen[5] Mannigfaltigkeiten

[3] Ein Homöomorphismus ist eine bijektive, stetige Abbildung, deren Umkehrabbildung ebenfalls stetig ist.

[4] Ein Diffeomorphismus ist immer auch ein Homöomorphismus, die Umkehrung gilt jedoch nicht.

[5] Eine riemannsche Mannigfaltigkeit heißt geodätisch konvex, wenn die global längenminimierende Kurve zwischen zwei beliebigen Punkten der Mannigfaltigkeit eine Geodäte dieser Mannigfaltigkeit ist.

die Abstände der Originaldaten nur unzureichend bzw. überhaupt nicht erhalten können [van07]. Lokale Verfahren liefern für solche Mannigfaltigkeiten bessere Ergebnisse und sind im Allgemeinen für ein breiteres Spektrum an Mannigfaltigkeiten geeignet. Werden Personen durch Kovarianzdeskriptoren repräsentiert — unabhängig davon, ob Einzelbilder von unterschiedlichen Personen oder eine Bildsequenz von einer beispielsweise spazieren gehenden Person betrachtet werden — liegen die Deskriptoren nicht zwangsweise auf geodätisch konvexen Untermannigfaltigkeiten im Sym_n^+ und können stark gekrümmt sein. Außerdem sind lokale Verfahren robuster gegenüber Ausreißern und Rauschen [Bel03]. Aus diesen Gründen wird in dieser Arbeit ein lokales Verfahren verwendet.

Unter den lokalen Ansätzen sind LLE und LE zwei der bekanntesten Verfahren (vgl. [Lee07]). Diese Ansätze benötigen keine Annahmen über die Form und Topologie einer Mannigfaltigkeit, weshalb sie für ein breites Anwendungsspektrum geeignet und somit auch für diese Arbeit interessant sind. Für diese Arbeit wurde LE gewählt, weil sich beim LLE die Optimierung der Verfahrensparameter schwierig gestaltet [Lee07].

Laplacian Eigenmaps

In diesem Abschnitt wird das lokale MaL Verfahren LE kurz vorgestellt, das in dieser Arbeit verwendet wird. Die Ausführung lehnt sich dabei an die Beschreibungen in [Bel03] und [Lee07] an. Gemäß [Bel03] gliedert sich das Verfahren in vier wesentliche Schritte:

1. symmetrische k-Nächste-Nachbarn-Suche zum Aufbau eines Nachbarschaftsgraphen,
2. Gewichtung der Kanten des Graphen,
3. Aufstellung der entsprechenden Laplace-Matrix des Graphen und
4. Lösung des daraus resultierenden generalisierten Eigenwertproblems.

Sei $\mathcal{X} = \{\boldsymbol{x}_i\}$, $i = 1, \ldots, n$, eine Menge von Punkten im $\mathbb{R}^D$ (Eingabepunkte), die auf (oder nahe bei) einer Mannigfaltigkeit liegen, für die eine niedrigdimensionale Repräsentation gefunden werden soll. Im ersten Schritt

wird eine k-Nächste-Nachbarn-Suche durchgeführt: Für jeden Eingabepunkt $\boldsymbol{x}_i \in \mathcal{X}$ werden seine k nächsten Nachbarn $N_i = \{\boldsymbol{x}_j\}$, $j = 1, \ldots, k$, bestimmt, wobei $\boldsymbol{x}_i \in N_i$ ist. Das Ergebnis wird anschließend durch einen ungerichteten Graphen repräsentiert, dessen Knoten die Punkte $\boldsymbol{x}_i$ sind und bei dem eine ungerichtete Kante zwischen zwei Knoten $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$ eingefügt wird, wenn $\boldsymbol{x}_i \in N_j$ oder $\boldsymbol{x}_j \in N_i$ ist, so dass eine symmetrische Nachbarschaftsbeziehung entsteht.

Das Ziel von LE (bzw. von MaL im Allgemeinen) ist die Abbildung von $\mathcal{X}$ auf eine Menge Punkte eines niedrigdimensionalen Raums, in diesem Fall auf die Punkte $\mathcal{Y} = \{\boldsymbol{y}_i\}$, $i = 1, \ldots, n$, $\boldsymbol{y}_i \in \mathbb{R}^d$ $(d \ll D)$, so dass die lokalen Abstände möglichst gut erhalten bleiben. Dazu wird aus dem Graphen eine gewichtete, symmetrische Adjazenzmatrix $\boldsymbol{W} = (w_{i,j})$ bestimmt, die sich aus folgender Gleichung ergibt:

$$w_{i,j} := \begin{cases} \exp\left(-\frac{\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_j)^2}{2T^2}\right), & \text{wenn } \boldsymbol{x}_i \in N_j\,, \\ 0\,, & \text{sonst.} \end{cases} \tag{4.17}$$

Die Kantengewichte werden dabei mit dem Wärmeleitungskern bestimmt, dessen funktionale Form einer Normalverteilung gleicht (wird in [Lee07] empfohlen). Dabei ist $\mathrm{d}(\cdot)$ der euklidische Abstand und T ein Skalierungsparameter, welcher der Temperatur in der Wärmeleitungsgleichung entspricht (vgl. [Gri09]).

Anschließend wird eine Kostenfunktion unüberwacht minimiert, welche fordert, dass *ähnliche* Eingabepunkte im hochdimensionalen Raum durch nahe beieinanderliegende Punkte im niedrigdimensionalen Raum repräsentiert werden. Je größer das Kantengewicht zwischen zwei benachbarten Eingabepunkten ist, desto größer ist die Ähnlichkeit. Dies entspricht dem Minimierungsproblem

$$\min \quad \sum_{i,j=1}^{n} w_{i,j} \cdot \mathrm{d}(\boldsymbol{y}_i, \boldsymbol{y}_j)^2\,. \tag{4.18}$$

Da $\boldsymbol{W}$ symmetrisch ist, lässt sich das Minimierungsproblem zu

$$\min \mathrm{Spur}(\boldsymbol{Y}\boldsymbol{L}\boldsymbol{Y}^T) \tag{4.19}$$

reduzieren (siehe [Lee07] für den Beweis der Äquivalenz dieser Gleichungen). $\boldsymbol{Y}$ ist eine Matrix, deren Spalten die Koordinaten der Punkte aus $\mathcal{Y}$ enthält und $\boldsymbol{L}$ entspricht der gewichteten *Laplace-Matrix* des Graphen und ist mit der Knotengrad-Matrix $\boldsymbol{D}$ gegeben durch

$$\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{W} \,. \tag{4.20}$$

Die Knotengrad-Matrix ist definiert durch

$$\boldsymbol{D} = (d_{i,j}) \,, \quad \text{mit } d_{i,i} := \sum_{j=1}^{n} w_{i,j} \tag{4.21}$$

und $d_{i,j} := 0$ für $i \neq j$.

Um die triviale Lösung $\boldsymbol{y}_1 = \boldsymbol{y}_2 = \cdots = \boldsymbol{y}_n = \vec{0}$ zu vermeiden, erfolgt die Minimierung unter der Nebenbedingung

$$\boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} = \boldsymbol{E}_{n \times n} \,, \tag{4.22}$$

wobei $\boldsymbol{E}_{n \times n}$ die $n \times n$ Einheitsmatrix ist.

Aus der Gleichung (4.20) ergibt sich eine diagonaldominante Laplace-Matrix, woraus weiter folgt, dass $\boldsymbol{L}$ positiv semidefinit ist und die Vektoren, welche die Gleichung (4.19) minimieren, den Eigenvektoren $\boldsymbol{V} = (\boldsymbol{v_1}, \ldots, \boldsymbol{v_n})$ entsprechen, die sich aus dem generalisierten Eigenwertproblem

$$\boldsymbol{L}\boldsymbol{V} = \boldsymbol{D}\boldsymbol{V}\boldsymbol{\Lambda} \tag{4.23}$$

ergeben (vgl. [Lee07]). Die Matrix $\boldsymbol{\Lambda}$ ist eine Diagonalmatrix mit den Eigenwerten.

Die Äquivalenz der Gleichungen (4.18) und (4.23) folgt wiederum aus dem *Rayleigh-Ritz-Theorem* [Hor85] (vgl. auch [Lee07]).

Der kleinste daraus resultierende Eigenwert λ_0 ist 0 und der zugehörige konstante Eigenvektor $\boldsymbol{v}_0$ ist $\vec{1}$. Die *beste* d-dimensionale Repräsentation $\boldsymbol{Y}$ ist durch die Eigenvektoren $\boldsymbol{v}_i$ gegeben, die zu den d kleinsten Eigenwerten $\lambda_i \neq 0 \leq \cdots \leq \lambda_{i+d}$ gehören:

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)^T \,. \tag{4.24}$$

Für den Fall, dass der Graph nicht zusammenhängend ist, also die Matrix $\mathbf{W}$ zu einer Blockdiagonalmatrix umgeordnet werden kann, muss das Lösen des generalisierten Eigenwertproblems (Gleichung (4.23)) für jeden Block, d.h. für jeden zusammenhängenden Untergraphen, separat durchgeführt werden.

Genauere und auch weiterführende Details bzgl. der Laplace-Matrix sind in [Chu95, vL07] zu finden. In der folgenden Aufzählung werden alternative Möglichkeiten und Berechnungen für die Teilschritte des LE aufgeführt.

Alternativen:

- Bei der k-Nächste-Nachbarn-Suche kann der Graph anstatt über die k nächsten Nachbarn alternativ über die Punkte aufgebaut werden, die einen Abstand kleiner $\epsilon > 0$ zu $\boldsymbol{x}_i$ haben, wobei hier direkt gilt, dass $\boldsymbol{x}_i \in N_j$ ist genau dann wenn $\boldsymbol{x}_j \in N_i$ ist.

- Anstatt die Kanten mittels des Wärmeleitungskerns zu gewichten, kann die Berechnung der Matrix $\boldsymbol{W}$ wahlweise durch folgende einfachere Gleichung erfolgen:

$$w_{i,j} = \begin{cases} 1\,, & \text{wenn } \boldsymbol{x}_i \in N_j\,, \\ 0\,, & \text{sonst.} \end{cases} \tag{4.25}$$

- Anstelle einer (generalisierten) Eigenwertzerlegung kann das Minimierungsproblem in (4.18) beispielsweise mit dem Verfahren der *Lagrange-Multiplikatoren* gelöst werden [Cam08].

- Eine äquivalente niedrigdimensionale Repräsentation zu der Repräsentation, die durch die Gleichung (4.24) gegeben ist, erhält man durch die Bestimmung der Eigenvektoren der normalisierten Laplace-Matrix $\boldsymbol{L}'$ [Ben04, Lee07]:

$$\boldsymbol{L}' = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{L} \boldsymbol{D}^{-\frac{1}{2}}\,. \tag{4.26}$$

  Damit muss anstelle des generalisierten Eigenwertproblems in Gleichung (4.23) nur eine einfache Eigenwertzerlegung durchgeführt wer-

den. Dadurch erhält man bis auf eine komponentenweise Skalierung eine äquivalente Repräsentation [Lee07]:

$$\boldsymbol{Y}^T = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \boldsymbol{D}^{\frac{1}{2}} (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)^T . \tag{4.27}$$

Dieser Weg wird auch in Kapitel 5 Anwendung finden.

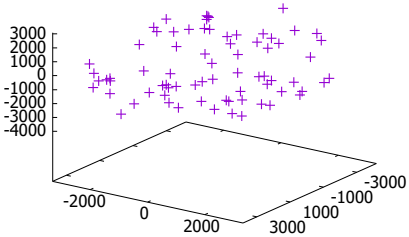
4.2.2 Untermannigfaltigkeiten im Sym_n^+

Eine Bildsequenz, die beispielsweise eine Aktivität einer Person oder ein Objekt in verschiedenen Ansichten zeigt, bildet — falls die Differenzen zwischen jeweils zwei zeitlich aufeinanderfolgenden Einzelbildern nicht zu groß sind — eine Mannigfaltigkeit (Untermannigfaltigkeit im euklidischen Raum) [Wei04]. Bei Bildsequenzen, die Personen abbilden, muss in der Regel von nichtlinearen Mannigfaltigkeiten ausgegangen werden. Abbildung 4.4a illustriert dies anhand eines Personen-Tracklets. Dabei wurden sowohl die Einzelbilder des Tracklets als auch die Kovarianzdeskriptoren, die aus den Einzelbildern berechnet wurden, mittels Hauptkomponentenanalyse und LE auf 3 reduziert. Die Bilder wurden dabei durch Vektoren mit der Dimension 8192 repräsentiert (Bildauflösung: 64 Pixel (Breite) × 128 Pixel (Höhe)). Die Kovarianzdeskriptoren wurden aus x-Koordinaten und RGB-Farbwerten berechnet.

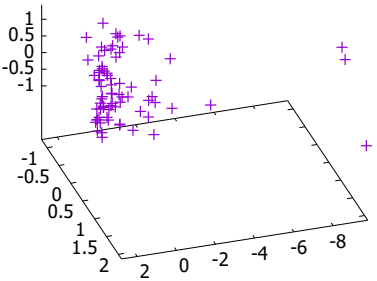
Das Ergebnis der linearen Dimensionsreduktion der Bildsequenz ist in Abbildung 4.4b zu sehen, wobei die Hauptkomponentenanalyse auf den als Vektoren repräsentierten Einzelbildern angewandt wurde. Das entsprechende nichtlineare Ergebnis, das mit LE berechnet wurde, ist in Abbildung 4.4d dargestellt.

Unter der Voraussetzung, dass die vektoriellen Repräsentanten der Einzelbilder eine Mannigfaltigkeit bilden, wird bei der Personendetektion und -wiedererkennung im Rahmen dieser Arbeit die Annahme getroffen, dass Kovarianzdeskriptoren, die Einzelbilder einer Bildsequenz repräsentieren, eine — in der Regel nichtlineare — Untermannigfaltigkeit im Sym_n^+ bilden. Diese Annahme wird durch die Abbildung 4.4e bekräftigt. Der Vollständigkeit halber ist in Abbildung 4.4c das Ergebnis der Hauptkomponentenanalyse zu sehen, die auch auf die Kovarianzdeskriptoren angewandt

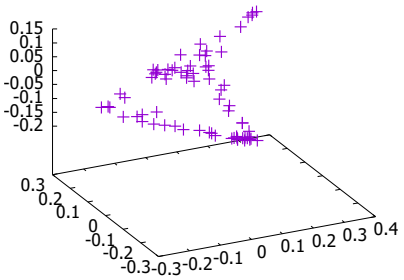
(a) Bildsequenz



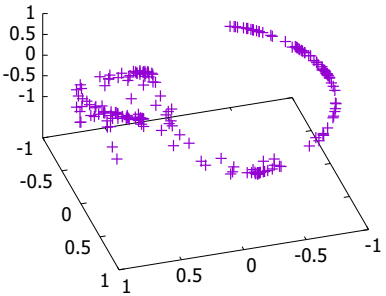
(b) Hauptkomponentenanalyse (Bildsequenz)



(c) Hauptkomponentenanalyse (Kovarianzdeskriptoren)



(d) LE (Bildsequenz)



(e) LE (Kovarianzdeskriptoren)

Abbildung 4.4: Eingabebildsequenz **(a)**, Dimensionsreduktion der Bildsequenz mittels Hauptkomponentenanalyse **(b)**, Dimensionsreduktion der aus den einzelnen Bildern berechneten Kovarianzdeskriptoren mittels Hauptkomponentenanalyse **(c)** sowie Dimensionsreduktion der Bildsequenz mittels LE **(d)** und Dimensionsreduktion der Kovarianzdeskriptoren mittels LE **(e)**.

wurde. In Abbildung 4.4e ist deutlich zu erkennen, dass die Punkte teilweise sehr dicht beieinander und auf stetigen Kurvensegmenten liegen, was in vielen Bildauswerte- und Bildverarbeitungsverfahren vorteilhaft genutzt werden kann. Ein naheliegendes Beispiel ist die Interpolation zwischen Bildern auf Pixelebene (z.B. zwischen Vektoren, die jeweils Grauwerte eines Bilds repräsentieren), die im niedrigdimensionalen Raum — zwischen den niedrigdimensionalen Repräsentanten — wesentlich effektiver als im hochdimensionalen Raum, in dem die Bilder in der Regel deutlich weiter auseinander liegen, durchgeführt werden kann.

Durch die Betrachtung von Kovarianzdeskriptor-Untermannigfaltigkeiten im Vergleich zu Mannigfaltigkeiten von Bildern soll die Diskriminanz zwischen Bildsequenzen unterschiedlicher Personen erhöht werden (vgl. Kapitel 7). Außerdem wird in Kapitel 5 die Annahme von Kovarianzdeskriptor-Untermannigfaltigkeiten für verschiedene Ansichten von Körperteilen getroffen, um die Klassifikation von Körperteilen zu verbessern.

5

Personendetektion anhand von Körperteilen in Einzelbildern

In diesem Kapitel werden die theoretischen Ansätze des mathematischen Rahmenwerks in ein Verfahren zur Personendetektion in Einzelbildern umgesetzt [Met10]. Das Ziel für die Detektion ist es, Körperteile — primär Kopf-Schulter-Partien — in Wärmebildern zu entdecken, wobei der Fokus auf der Klassifikation gegebener Bildregionen liegt, die durch beispielsweise Schwellwertverfahren erzeugt wurden. Für die Repräsentation der Körperteile wird die riemannsche Mannigfaltigkeit der Kovarianzdeskriptoren verwendet, die in Abschnitt 4.1 beschrieben ist. Außerdem wird der LE Algorithmus — ähnlich wie in [Jia09] — zu einem überwachten Ansatz für eine diskriminativere Repräsentation von Kovarianzdeskriptoren erweitert. Die Abbildung zwischen den Kovarianzdeskriptoren und der Einbettung, die durch LE nur implizit gelernt wird, wird mittels dem *radiale Basisfunktionen* (RBF)-Interpolations-Rahmenwerk in [Elg08] gelernt.

Anhand des vorgestellten Verfahrens, das auf Kovarianzdeskriptoren beruht, werden im Anwendungsgebiet der Körperteildetektion die Ergebnisse in [Tuz07] und [Pai07] bekräftigt, dass mit Kovarianzdeskriptoren bessere Detektionsergebnisse als mit der direkten Verwendung der zugrunde liegenden Merkmale erzielt werden können. Zudem wird die Annahme aus Abschnitt

4.2.2, dass Kovarianzdeskriptoren, die Einzelbilder eines Tracklets repräsentieren, eine Untermannigfaltigkeit im Sym_n^+ bilden, gewinnbringend für Körperteilklassen angewandt. Dazu werden mittels überwachtem MaL Mannigfaltigkeiten einzelner Körperteile gelernt, um die Intra-Variationen einer Körperteilklasse zu minimieren und die Inter-Variationen zwischen unterschiedlichen Körperteilmannigfaltigkeiten bzw. zu negativen Proben zu maximieren.

Das Personendetektionsverfahren soll in der Fahrzeugumfelderfassung — auf mobilen Plattformen — einsetzbar sein und primär zur militärischen Gefahrenabwehr zur Anwendung kommen, in denen Personen aus einem fahrenden Fahrzeug in urbanen Gebieten erkannt werden sollen. Da die Besatzung des Fahrzeugs nicht immer direkten Sichtkontakt mit der Umgebung hat, ist die automatische bildbasierte Detektion von Personen sowohl für die Erhöhung der Verkehrssicherheit als auch für die Erkennung von Bedrohungen eine hilfreiche Unterstützung. Automatische Assistenzsysteme zur frühzeitigen Detektion aller Personen in der Szene, auch wenn sie fast vollständig verdeckt sind, sind deshalb eine enorme Unterstützung. Es ist folglich wichtig, dass solche Systeme die Personen anhand von Körperteilen — insbesondere Kopf-Schulter-Partien — erkennen können, und dass zudem unabhängig davon, ob sich die Person gerade bewegt oder ob sie stillsteht. Dies bietet den zusätzlichen Vorteil, dass auch Einzelpersonen in Menschenmengen entdeckt werden können.

Verfahren zur Körperteildetektion mittels handentworfener Merkmale lassen sich in direkte und indirekte Ansätze unterteilen. Indirekte Verfahren werden oft für die Detektion in Bildfolgen von unbewegten Kameras verwendet [Lee04]. Bei diesen Ansätzen werden zunächst bewegte Objekte vom Hintergrund getrennt, die im zweiten Schritt als ein (bestimmtes) Körperteil oder *kein Körperteil* klassifiziert werden. Indirekte Verfahren können zwar auch für Bildsequenzen von Kameras auf mobilen Plattformen verwendet werden, wofür jedoch Vorverarbeitungsalgorithmen notwendig werden. Die einzelnen Bilder müssen dafür zunächst aufeinander registriert werden, um Bewegungen detektieren zu können, die sich von der Eigenbewegung der mobilen Plattform unterscheiden. Da in diesem Kapitel zugrunde liegende Anwendungsszenario auch stillstehende Personen detektiert werden sollen, sind für diese Arbeit direkte Verfahren besser geeignet. Sie detektieren Personen in Einzelbildern beispielsweise anhand ihrer Hautfarbe oder Kontur.

In [Vio03] wird einer der ersten beliebten direkten Ansätze vorgestellt. Das Verfahren verwendet unterschiedliche Rechteckfilter für die Erscheinung und Bewegungsmuster, welche mittels einer Kaskade *einfacher* Klassifikatoren klassifiziert werden. Der Ansatz basiert auf den einfachen Rechteckfiltern für die Gesichtserkennung, die in [Vio01] vorgestellt werden. Darüber hinaus gibt es im Bereich der Fahrerassistenz zwei weitere weit verbreitete direkte Ansätze, die sich für die Detektion von Fußgängern bewährt haben: ein auf sowohl Konturen als auch Texturen basierter Ansatz [Gav07] und ein Verfahren, das *Implicit Shape Models* (ISM) verwendet [Lei04]. Seit der Veröffentlichung des ISM wurden zudem zahlreiche Erweiterungen dafür vorgestellt. In [Bra12, Bra14] wird beispielsweise eine neue ISM-Voting-Strategie vorgeschlagen, um anatomische Landmarken von Personen präzise lokalisieren zu können, worüber sich auch Körperteile robust detektieren lassen. Ein anderer sehr weit verbreiteter konturbasierter Ansatz ist in [Dal05] veröffentlicht. Das darin beschriebene Verfahren klassifiziert mittels einer SVM dicht abgetastete Histogramme von orientierten Gradienten (HOG). Es wurde von dem populären SIFT-Ansatz in [Low04] motiviert und zeichnet sich durch gute Ergebnisse bei der Personendetektion aus.

Im Folgenden wird zunächst ein Verfahren zur Detektion von Körperteilen in Wärmebildern mittels HOG vorgestellt, das als Referenzverfahren verwendet wird. Das Detektionsverfahren entspricht dem Ansatz in [Dal05] mit einer zusätzlichen Suche von zusammenhängenden Bildregionen mit hohen Grauwerten vor der eigentlichen Detektion und die Verwendung von Körperteil- anstatt Ganzkörper-Konturen. Der Ansatz [Dal05] beruht auf handentworfenen Merkmalen und war seiner Zeit aufgrund guter Detektionsergebnisse Stand der Forschung im Gebiet der Personendetektion in Einzelbildern. In der eigenen Arbeit [Met07] konnten mit HOG gute Ergebnisse bei der Detektion von Kopf-Schulter-Partien in Wärmebildern erzielt werden.

Anschließend wird das darauf aufbauende Detektionsverfahren vorgestellt, das Körperteile anhand von Körperteilmannigfaltigkeiten detektiert [Met10]. Für die Personenrepräsentation werden Kovarianzdeskriptoren verwendet, die aus Gradientenmerkmalen (HOG-ähnliche Merkmale) berechnet werden. Der Detektionsschritt basiert auf einem überwachten MaL Algorithmus, welcher Mannigfaltigkeiten einzelner Körperteile lernt und zwar so, dass die Intra-Variationen einer Körperteilmannigfaltigkeit (Intra-Diskriminanz von Bildern einer Körperteilklasse) klein und die

Abbildung 5.1: Beispielbilder aus dem Wärmebilddatensatz (Training). Der Datensatz besteht aus Sequenzen von Wärmebildern, die während verschiedenen Ortsdurchfahrten eines kameratragenden Fahrzeugs durch ein Truppenübungsdorf entstanden.

Inter-Variationen (Inter-Diskriminanz) zwischen unterschiedlichen Körperteilmannigfaltigkeiten bzw. zu negativen Proben groß wird. Anhand eines Wärmebilddatensatzes (siehe Abbildung 5.1), der in einem Truppenübungsdorf im Rahmen einer Bundeswehrübung akquiriert wurde, wird gezeigt, dass der hier vorgeschlagene MaL Ansatz den Klassifikationsschritt von handentworfenen Merkmalen für Detektionssysteme verbessern kann.

Überwachtes Manifold Learning. Seit Anfang 2000 wurden überwachte MaL Algorithmen in der Mustererkennung aufgrund der guten Ergebnisse immer beliebter [Kou03, dR03, Bel04, Lia05, Wan05, Ten07]. Überwachtes MaL unterscheidet sich von den unüberwachten Algorithmen in der Art der Projektion der Eingabedaten, die schon klassifiziert vorliegen. Die überwachten Verfahren projizieren Daten derselben Klasse so, dass sie nahe beieinander liegen und Daten unterschiedlicher Klassen so, dass sie weit voneinander entfernt liegen, bei gleichzeitiger Berücksichtigung der intrinsischen Geometrie der einzelnen Mannigfaltigkeiten. Die intrinsische Geometrie soll dabei möglichst genau rekonstruiert werden. Der Zweck der Dimensionsreduktion ist somit nicht nur die Reduktion der Dimension der vorhandenen Eingabedaten, sondern auch die Erhöhung des Abstands zwischen Daten unterschiedlicher Klassen. Außerdem ermöglicht eine überwachte Dimensionsreduktion das Auffinden der zugehörigen Mannigfaltigkeit neuer Punkte, was bei Klassifikationsverfahren genutzt werden kann. Eine ausführliche theoretische Analyse von überwachten MaL Verfahren bezüglich Klassifikationsleistung ist in [Vur15] zu finden.

5.1 HOG-basierte Körperteildetektion

Das HOG-basierte Verfahren von Dalal et al. [Dal05] wurde im Rahmen dieser Arbeit für die Detektion von Körperteilen in Wärmebildern angepasst und trainiert. Das verfolgte Anwendungsziel ist die zuverlässige einzelbildbasierte Detektion von Personen im urbanen Gebiet anhand von insbesondere Kopf-Schulter-Partien. Im Folgenden werden die einzelnen Verfahrensschritte und -phasen kurz beschrieben: Zunächst wird die Berechnung der HOG-Deskriptoren beschrieben, gefolgt vom Training eines

Abbildung 5.2: Gefaltete Bildausschnitte $\boldsymbol{I}_{\boldsymbol{R}_x}$ und $\boldsymbol{I}_{\boldsymbol{R}_y}$.

Klassifikators und schließlich die Onlinephase, in der zur Laufzeit die Körperteile detektiert werden.

5.1.1 HOG-Deskriptor-Berechnung

Die HOG-Deskriptor-Berechnung erfolgt für einzelne Bildausschnitte $\boldsymbol{I}_{\boldsymbol{R}}$ (Ausschnitte aus Intensitätsbildern), die zunächst auf eine einheitliche (vorgegebene) Größe skaliert werden. Die Berechnung besteht aus vier wesentlichen Schritten: Faltung der Bilder, um Gradienten-Bilder zu erhalten, Erstellung von Gradienten-Histogrammen für Bildzellen, Zusammenfassung von Zellen- zu Blockdiagrammen und Konkatenation der Blockdiagramme.

Im ersten Verfahrensschritt werden mittels den Filterkernen $[-1, 0, 1]$ und $[-1, 0, 1]^T$ Bildgradienten (x- und y-Richtung) innerhalb der skalierten Bildausschnitte berechnet. Die resultierenden Bilder werden als $\boldsymbol{I}_{\boldsymbol{R}_x}$ und $\boldsymbol{I}_{\boldsymbol{R}_y}$ bezeichnet (vgl. Abbildung 5.2).

In Schritt 2 werden aus $\boldsymbol{I}_{\boldsymbol{R}_x}$ und $\boldsymbol{I}_{\boldsymbol{R}_y}$ Histogramme berechnet. Dafür werden die Richtungen $\phi(x, y)$ und Magnituden $m(x, y)$ der Gradienten für

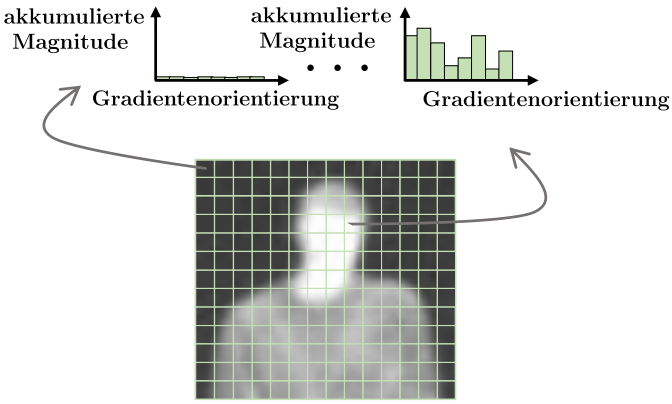


Abbildung 5.3: Berechnung der Zellenhistogramme: Ein skalierter Bildausschnitt wird in gleich große Zellen geteilt und für jede Zelle ein Zellenhistogramm aus den Gradientenrichtungen und -magnituden berechnet.

jedes Pixel (x, y) des Bildausschnitts $\boldsymbol{I}_{\boldsymbol{R}}$ berechnet, die sich aus folgenden Gleichungen ergeben:

$$\phi(x,y) = \tan^{-1}\left(\frac{\boldsymbol{I}_{\boldsymbol{R}_y}(x,y)}{\boldsymbol{I}_{\boldsymbol{R}_x}(x,y)}\right), \tag{5.1}$$

$$m(x,y) = \sqrt{\left(\boldsymbol{I}_{\boldsymbol{R}_x}(x,y)^2 + \boldsymbol{I}_{\boldsymbol{R}_y}(x,y)^2\right)}. \tag{5.2}$$

Außerdem wird der Bildausschnitt in Zellen aufgeteilt und für jede Zelle ein Histogramm (Zellenhistogramm) aus den Gradientenrichtungen bestimmt. Ein Zellenhistogramm wird in acht gleich breite Intervalle aufgeteilt, in die die Gradienten anhand ihrer Richtung einsortiert werden. Zusätzlich werden die Gradienten mit ihrer Magnitude gewichtet und pro Zelle aufsummiert. Das Vorzeichen der Gradienten wird dabei nicht betrachtet. Damit beinhaltet ein Zellenhistogramm nur Gradientenrichtungen zwischen $0°$ und $180°$. Um Quantisierungsfehler zu mindern, werden die Magnituden außerdem bilinear zwischen benachbarten Intervallen interpoliert. Abbildung 5.3 illustriert die Berechnung der Zellenhistogramme.

Die Verwendung vorzeichenloser Gradienten, die Aufteilung des Histogramms in acht Intervalle sowie die Gewichtung mit der Magnitude

wurden mittels Parameteroptimierung ermittelt. Bei der Gesamtkörper-Personendetektion in [Dal05] erzielten Dalal und Triggs die besten Ergebnisse mit ähnlichen Zellhistogrammen, nur die Anzahl der Intervalle unterscheidet sich: Ihre Optimierung für die Personendetektion anhand des gesamten Körpers ergab eine optimale Aufteilung des Histogramms in neun Intervalle.

In Schritt 3 werden benachbarte Zellenhistogramme zu Blockdiagrammen zusammengefasst (vgl. Abbildung 5.4). Dies erfolgt durch Konkatenation der Zellenhistogramme innerhalb eines Blocks, beginnend mit dem Histogramm der linken oberen Zelle (zeilenweise Abarbeitung). Für die Körperteildetektion werden quadratische Blöcke verwendet, die 2×2 Zellen mit einer Zellengröße von 8×8 Pixeln beinhalten. Es können auch rechteckige oder runde Blöcke betrachtet werden. Bei der Körperteildetektion konnten die besten Ergebnisse mit den gerade genannten quadratischen Blöcken erzielt werden.

Die Blockdiagramme werden außerdem normalisiert, um eine hohe Robustheit gegenüber Änderungen in der Beleuchtung und des Vorder-Hintergrundkontrasts zu erhalten. Dies erfolgt durch die Normierung der Blockdiagrammvektoren mittels der euklidischen Norm. Es muss beachtet werden, dass dominante Gradientenrichtungen in den Blockdiagrammen das Detektionsergebnis möglicherweise verschlechtern können. Eine Möglichkeit dies zu verhindern ist die Begrenzung der Werte der Elemente des Blockdiagrammvektors. Im Rahmen dieser Arbeit beeinflusste die Hysterese das Detektionsergebnis nicht.

Im letzten Schritt wird der Block um eine Zelle weitergeschoben, so dass die meisten Zellen in mehreren Blockdiagrammen berücksichtigt werden. Diese Überlappung bewirkt eine Art gleitende lokale Kontrastanpassung und verbessert das Detektionsergebnis.

Der HOG-Deskriptor ergibt sich schließlich durch Konkatenation der Blockhistogramme, beginnend mit dem Block links oben (zeilenweise Abarbeitung).

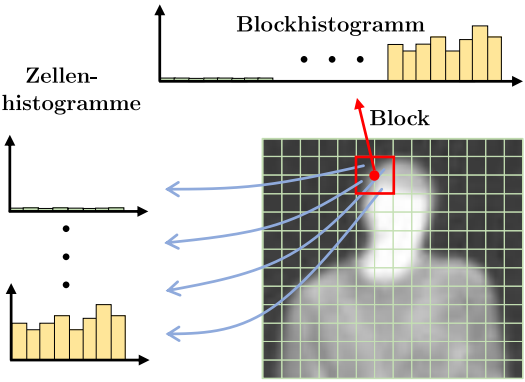
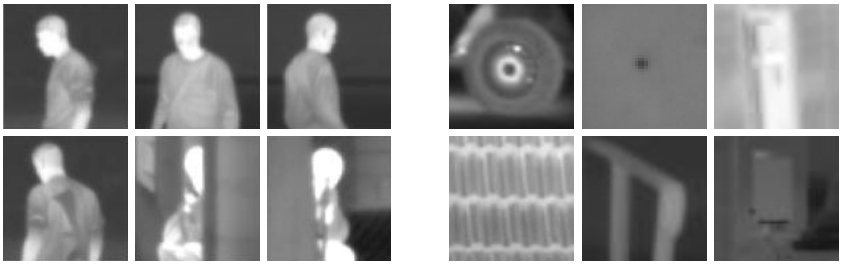


Abbildung 5.4: Berechnung eines Blockdiagramms: Die Zellenhistogramme innerhalb eines Blocks werden durch Konkatenation zu einem Blockdiagramm zusammengefasst.

5.1.2 Trainingsphase

In der Trainingsphase werden die HOG-Deskriptoren zunächst für annotierte Bildausschnitte von sowohl Körperteilpartien (positive Beispiele) als auch für andere Bildausschnitte (negative Beispiele), wie sie exemplarisch in Abbildung 5.5 dargestellt sind, berechnet. Alle Bildausschnitte werden dafür auf die Größe von 64×64 Pixel skaliert.

Anschließend wird ein Klassifikator auf den Deskriptoren trainiert. Ein Klassifikator, der dafür oft eingesetzt wird, ist die SVM, die z.B. auch in [Met07] für die Kopf-Schulter-Detektion verwendet wird. Für die Evaluation in dieser Arbeit wird bedingt durch eine kleine Trainingsmenge ein k-Nächste-Nachbarn-Klassifikator verwendet. Während in [Met07] für die Auswertung des Verfahrens eine SVM mit 1150 Kopf-Schulter-Proben und 1150 negativen Beispielen trainiert wurden, werden im Rahmen der Körperteildetektion pro Klasse lediglich 250 Proben betrachtet.



(a) Positive Beispiele **(b)** Negative Beispiele

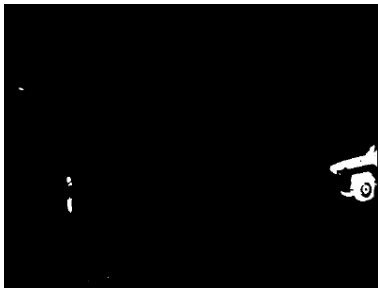
Abbildung 5.5: Beispiele von Bildausschnitten, die für das Training der HOG-basierten Körperteildetektion verwendet wurden. Als positive Beispiele wurden Körperteilpartien verwendet, von denen einige teilweise verdeckt sind **(a)**. Die negativen Beispiele wurden zufällig aus dem MOUT-Datensatz extrahiert **(b)**.

5.1.3 Onlinephase

In der Onlinephase, also zur Laufzeit, werden im ersten Schritt Bildregionen gesucht, die eine Körperteilpartie beinhalten können. Im Allgemeinen sind das Bereiche mit hohen Grauwerten, da das Verfahren auf Wärmebildern arbeitet (vgl. Abbildung 5.6). Dazu wird das gesamte Bild zunächst mittels eines Schwellwerts gefiltert. Daraus resultiert eine binäre Bildmaske mit derselben Auflösung wie das Eingangsbild, wobei die Pixelwerte der Maske auf 0 gesetzt werden, wenn die entsprechende Pixelposition im Eingangsbild einen Grauwert kleiner einem vorgegebenen Schwellwert aufweist, andernfalls auf 1. Damit muss der Körperteildetektor nicht das gesamte Bild nach Körperteilpartie absuchen, sondern nur die durch die Schwellwertfilterung ermittelten Bildregionen. Dadurch können deutlich kürzere Laufzeiten erzielt werden und die Anzahl falscher Detektionshypothesen ist in der Regel kleiner. Abgedeckte Körperteile, deren Wärmestrahlung dadurch unterdrückt wird, können dann allerdings nicht mehr detektiert werden.

Im zweiten Schritt wird im ursprünglichen Wärmebild (Abbildung 5.6a) ein quadratisches Detektionsfenster über die gefundenen Bildregionen im

(a) Beispiel aus dem Wärmebildatensatz



(b) Binäre Bildmaske nach der Schwellwertfilterung

Abbildung 5.6: Beispiel für ein mit einem Schwellwert gefiltertes Wärmebild.

Wärmebild geschoben (dieser Vorgang wird für verschiedene Skalierungen des Bilds bzw. Detektionsfensters wiederholt). Ist der Anteil der Pixel, die weiß sind, größer als 5%, wird ein HOG-Deskriptor berechnet und zur Klassifikation der SVM übergeben.

5.2 Manifold Learning basierte Körperteildetektion

In diesem Abschnitt wird das MaL basierte Verfahren zur Detektion von Körperteilen in Wärmebildern vorgestellt. Der Ansatz basiert auf Mannigfaltigkeiten von Körperteilen. Die Mannigfaltigkeiten einzelner Körperteile werden dabei überwacht gelernt, und zwar so, dass die Intra-Variationen (Intra-Diskriminanz von Bildern einer Körperteilklasse) einer Körperteilmannigfaltigkeit klein und die Inter-Variationen (Inter-Diskriminanz) zwischen unterschiedlichen Körperteilmannigfaltigkeiten bzw. zu negativen Proben groß wird (vgl. Abbildung 5.7). Dazu wurde der in Abschnitt 4.2.1 beschriebene MaL Algorithmus LE — ähnlich wie in [Jia09] — für überwachtes Lernen angepasst (vgl. [Met10]).

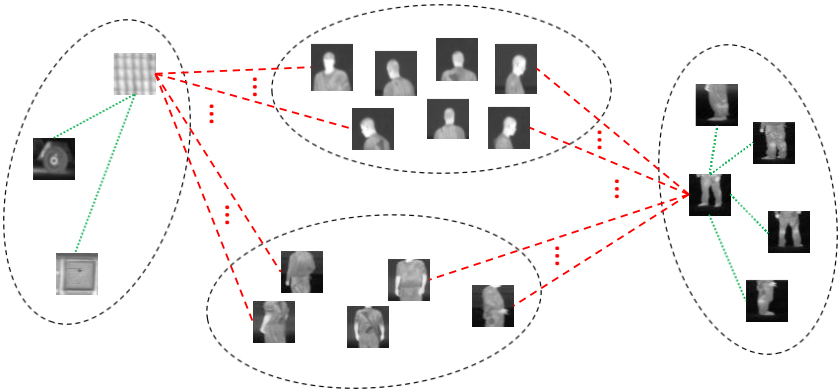


Abbildung 5.7: Verwendete Körperteilklassen (rechte drei Cluster) und eine Klasse mit negativen Beispielen (linkes Cluster), die beim Lernen der Mannigfaltigkeiten mitberücksichtigt wurden. Anhand zweier Beispiele sind die Intra- und Inter-Diskriminanzen schematisch dargestellt. Die Intra-Diskriminanzen für die beiden Beispielbilder — oberes Bild im linken Cluster und linkes Bild im rechten Cluster — sind durch die grün gepunkteten Linien und einige Inter-Diskriminanzen durch die rot gestrichelten Linien gekennzeichnet.

Der im Rahmen dieser Arbeit erarbeitete Ansatz zur Körperteildetektion verwendet Kovarianzdeskriptoren für die Repräsentation der Körperteile. Wie bereits in Abschnitt 3 angesprochen, können durch Kovarianzdeskriptoren, anstatt die zugrunde liegenden Merkmale direkt zu verwenden, höhere Detektionsraten erzielt werden. In [Tuz07, Pai07] beispielsweise wurde mit Kovarianzdeskriptoren zur Detektion von Personen oder Körperteilen bessere Ergebnisse als mit HOG-basierten Detektionsverfahren erzielt, trotz Verwendung sehr ähnlicher Merkmale, weswegen ab den Veröffentlichungen dieser Ergebnisse der Fokus auf die Kovarianzdeskriptoren für die weiteren Arbeiten im Rahmen der Personendetektion in niedrig aufgelösten Bildern gelegt wurde.

Die Kovarianzdeskriptoren für die MaL basierte Körperteildetektion basieren, wie auch die HOG, auf Gradientenmerkmalen. Der zugrundeliegende Merkmalsvektor $\boldsymbol{f}_{(x,y)}$ für die Koordinate (x, y) bzgl. der linken oberen Ecke eines Bildausschnitts $\boldsymbol{I_R}$ mit der Koordinate $(0, 0)$ ist definiert durch

$$\boldsymbol{f}_{(x,y)} = \begin{pmatrix} X(x,y) \\ Y(x,y) \\ \left|\frac{\delta \boldsymbol{I_R}(x,y)}{\delta x}\right| \\ \left|\frac{\delta \boldsymbol{I_R}(x,y)}{\delta y}\right| \\ \left|\frac{\delta^2 \boldsymbol{I_R}(x,y)}{\delta x^2}\right| \\ \left|\frac{\delta^2 \boldsymbol{I_R}(x,y)}{\delta y^2}\right| \\ \sqrt{\left(\boldsymbol{I_{R_x}}(x,y)^2 + \boldsymbol{I_{R_y}}(x,y)^2\right)} \\ \tan^{-1}\left(\left|\frac{\boldsymbol{I_{R_y}}(x,y)}{\boldsymbol{I_{R_x}}(x,y)}\right|\right) \end{pmatrix}, \tag{5.3}$$

mit $X(x, y) = x$ und $Y(x, y) = y$.

Sei b die Rechteckbreite und h die Rechteckhöhe von $\boldsymbol{I_R}$. Der Kovarianzdeskriptor $\boldsymbol{\Sigma_R}$ ist dann gegeben durch

$$\boldsymbol{\Sigma_R} = \frac{1}{b \cdot h} \sum_{x=0}^{b-1} \sum_{y=0}^{h-1} \left(\boldsymbol{f}_{(x,y)} - \boldsymbol{\mu_R}\right)\left(\boldsymbol{f}_{(x,y)} - \boldsymbol{\mu_R}\right)^T, \tag{5.4}$$

wobei $\boldsymbol{\mu_R}$ der Mittelwertvektor aus $\{\boldsymbol{f}_{(x,y)}\}$, $x, = 0, \ldots, b-1$, $y, = 0, \ldots, h-1$, ist (vgl. Abschnitt 3.4.1).

Bei dem hier vorgestellten Ansatz liegt die Annahme zugrunde, dass die Kovarianzdeskriptoren einer Körperteilklasse auf oder nahe bei einer (möglicherweise nichtlinearen) Mannigfaltigkeit liegen, die im Raum der positiv definiten Kovarianzdeskriptoren eingebettet ist (vgl. Abschnitt 4.2.2) und nicht neben einer Untermannigfaltigkeit eines anderen Körperteils liegt. Die Gesamtmannigfaltigkeit, die sich aus den einzelnen Körperteilmannigfaltigkeiten und den negativen Beispielen bestimmt, wird mit dem für überwachtes Lernen angepassten LE Algorithmus gelernt. Ein Lernverfahren, das auf einem lokalen Verfahren wie dem LE beruht, hat den Vorteil, dass es sich sehr gut für stark gekrümmte Mannigfaltigkeiten eignet, wie sie im Allgemeinen bei Körperteilen auftreten. Außerdem erzielen lokale Ansätze — im Vergleich zu globalen Methoden — für ein breiteres Spektrum an Mannigfaltigkeiten verwendbare Ergebnisse: z.B. für Mannigfaltigkeiten, deren globale Geometrie nicht euklidisch ist und deren lokale Geometrie der euklidischen ähnelt [Sil03].

5.2.1 Trainingsphase

In der Trainingsphase wird eine d-dimensionale Gesamtrepräsentation des Trainingsdatensatzes mittels überwachtem MaL gelernt. Es können sowohl Repräsentationen, die jeweils nur eine Körperteilklasse berücksichtigen, als auch Gesamtrepräsentationen für mehrere Körperteilklassen gelernt werden.

d-dimensionale Gesamtrepräsentation

Analog zur HOG-basierten Körperteildetektion werden zunächst Deskriptoren für annotierte Bildausschnitte berechnet, von sowohl Körperteilen (positive Beispiele) als auch für andere Bildausschnitte (negative Beispiele), wie sie exemplarisch in der Abbildung 5.6 dargestellt sind. Bei dem MaL basierten Detektionsansatz werden Kovarianzdeskriptoren verwendet. Die Kovarianzdeskriptoren werden für einheitlich groß skalierte Bildregionen berechnet, so dass sie — unter der Annahme, dass die Körperteile die

Bildregionen ausfüllen — skalierungsinvariant sind. Eine Rotationsinvarianz ist aufgrund der zugrundeliegenden Merkmale nicht gegeben, weshalb unterschiedlich rotierte Körperteile im Trainingsdatensatz mitberücksichtigt werden müssen. Sind keine entsprechende Trainingsdaten vorhanden, müssen die Kovarianzdeskriptoren rotationsinvariant konstruiert werden, indem z.B. keine örtliche Struktur betrachtet wird — was Farbhistogrammen ähnelt — oder die Kovarianzdeskriptoren aus rotationsinvarianten Merkmalen, wie z.B. SDALF, berechnet werden.

Sei $\mathcal{A} = \{\boldsymbol{\Sigma}_i\}$, $i = 1, \ldots, n$, die Trainingsmenge, bestehend aus n Kovarianzdeskriptoren von sowohl Körperteilen als auch von negativen Beispielen, sowie $\mathcal{L} = \{l_i\}$, $i = 1, \ldots, n$, die zugehörigen Klassenlabels, wobei $l_i \in \{1, \ldots, C\}$ (C =Anzahl der Klassen).

Für jeden Kovarianzdeskriptor $\boldsymbol{\Sigma}_i \in \mathcal{A}$ werden zunächst seine k nächsten Nachbarn im Sym_n^+ gesucht. Die k nächsten Nachbarn bestimmen sich ähnlich wie in [Jia09] durch das Ähnlichkeitsmaß

$$\mathrm{s}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j) = \begin{cases} 1 - \exp\left(-\mathrm{g}^2(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)\right), & \text{wenn } l_i = l_j\,, \\ \exp\left(\mathrm{g}^2(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)\right), & \text{sonst}\,, \end{cases} \tag{5.5}$$

wobei in dieser Gleichung $\mathrm{g}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)$ der geodätische Abstand zwischen $\boldsymbol{\Sigma}_i$ und $\boldsymbol{\Sigma}_j$ ist, der mit der affin-invarianten riemannschen Metrik in Abschnitt 4.1.1 berechnet wird. Die Anzahl der Nachbarn k sollte abhängig von der Trainingsdatenmengengröße gewählt werden: Je größer die Trainingsdatenmenge, desto höher sollte k gewählt werden.

Anschließend wird die normalisierte Laplacian Matrix bestimmt, die sich aus folgender Gleichung ergibt (vgl. Abschnitt 4.2.1):

$$\boldsymbol{L}' = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{L} \boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{D}^{-\frac{1}{2}} (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{D}^{-\frac{1}{2}}\,. \tag{5.6}$$

Die Matrix $\boldsymbol{W} = (w_{i,j})$ ist eine gewichtete, symmetrische Adjazenzmatrix auf Basis der Ähnlichkeitswerte, die sich berechnet durch

$$w_{i,j} := \begin{cases} \exp\left(-\frac{\mathrm{s}(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)}{2T^2}\right), & \text{wenn } \boldsymbol{\Sigma}_j \in N_i\,, \\ 0\,, & \text{sonst}\,, \end{cases} \tag{5.7}$$

wobei der Skalierungsparameter $T = 1$ ist. Die Matrix $\boldsymbol{D} = (d_{i,j})$ ist die Knotengrad-Matrix aus Gleichug (4.21), die sich aus

$$\boldsymbol{D} = (d_{i,j}), \quad \text{mit } d_{i,i} := \sum_{j \in N_i} w_{i,j} \tag{5.8}$$

ergibt.

Die niedrigdimensionale (d-dimensionale) Gesamtrepräsentation $\boldsymbol{Y}$ ergibt sich aus den Eigenvektoren $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ von $\boldsymbol{L}'$, die zu den kleinsten Eigenwerten $\lambda_i \neq 0 \leq \cdots \leq \lambda_{i+d}$ von $\boldsymbol{L}'$ gehören (vgl. Kapitel 4.2.1):

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \boldsymbol{D}^{\frac{1}{2}} (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)^T . \tag{5.9}$$

Abbildung 5.8 zeigt eine Übersicht über den ersten Teil der Trainingsphase.

Parameter d. Mittels überwachten MaL Methoden kann eine lineare Separierbarkeit der Daten erreicht werden [Vur15]. Vural et al. haben zudem gezeigt, dass bei einem 2-Klassen-Problem $d = 1$ dafür im Allgemeinen ausreicht. Bei mehreren Klassen ergibt sich die Dimension aus der Summe der Dimension der einzelnen Klassen (Einbettungen), die eine lineare Separierbarkeit innerhalb jeder Klasse sicherstellen (siehe [Vur15] für Details).

5.2.2 Onlinephase

In der Onlinephase der Körperteildetektion werden zunächst wie bei der HOG-basierten Körperteildetektion Bildregionen mit hohen Grauwerten gesucht. Dazu wird mittels Schwellwertverfahren eine binäre Bildmaske aus dem Wärmebild bestimmt, anhand dieser die Bildregionen bestimmt werden. Zudem wird mehrmals, in unterschiedlichen Skalierungen, ein Detektionsfenster über die gefundenen Bildregionen geschoben und jeweils ein Kovarianzdeskriptor für den entsprechenden Bildausschnitt im Wärmebild berechnet. Anschließend wird eine nichtlineare Einbettung zwischen der lokalen Umgebung der extrahierten Kovarianzdeskriptoren und der gelernten d-dimensionalen Mannigfaltigkeit bestimmt.

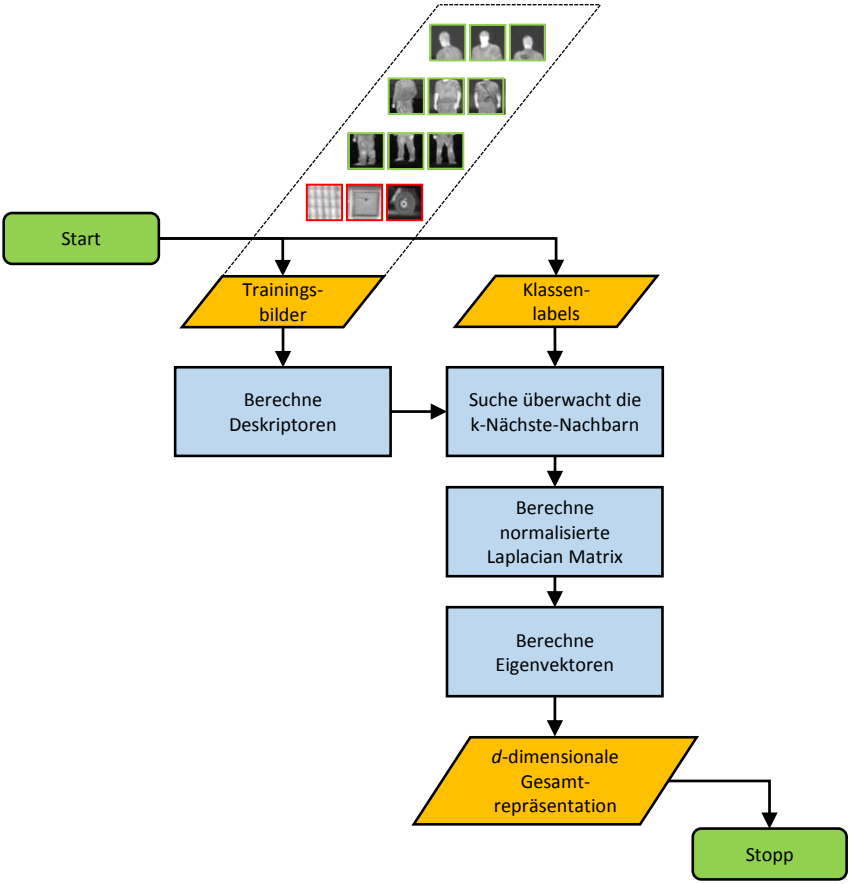


Abbildung 5.8: Übersichtsdiagramm über die Trainingsphase: Das Lernen der d-dimensionalen Gesamtrepräsentation des Trainingsdatensatzes. Mittels dem Ähnlichkeitsmaß, das durch die Gleichung (5.5) definiert ist, wird eine d-dimensionale Gesamtrepräsentation aller Untermannigfaltigkeiten überwacht gelernt.

Nichtlineare Einbettung

Sei $\mathcal{B} = \{\boldsymbol{\Theta}_i\}$, $i = 1, \ldots, n$ die Menge der zu klassifizierenden Kovarianzdeskriptoren. Die Klassifikation der Deskriptoren, also die Zuordnung der Deskriptoren zu einer Körperteilmannigfaltigkeit oder der *negativen* Klasse, erfolgt anhand von Merkmalen der niedrigdimensionalen (d-dimensionale) Gesamtrepräsentation $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ der Untermannigfaltigkeiten der Kovarianzdeskriptoren $\mathcal{A} = \{\boldsymbol{\Sigma}_i\}, i = 1, \ldots, n$ aus dem Training, die sowohl die unterschiedlichen Körperteile als auch die negativen Beispiele beinhalten sowie die Klassenzugehörigkeit der Trainings-Kovarianzdeskriptoren. Die Kovarianzdeskriptoren $\mathcal{B}$ müssen dazu auf die d-dimensionale Mannigfaltigkeit abgebildet werden. So wie die meistens anderen MaL Algorithmen lernt auch das LE Verfahren die Einbettung allerdings nur implizit.

Damit neue Kovarianzdeskriptoren auf die gelernte, im Raum der Kovarianzdeskriptoren eingebettete, d-dimensionale Mannigfaltigkeit projiziert werden können, muss die nichtlineare Einbettung gelernt werden, also der (lokale) Homöomorphismus zwischen der gelernten d-dimensionalen Mannigfaltigkeit und dem Raum der Kovarianzdeskriptoren. Das Lernen einer glatten (nichtlinearen) Abbildung anhand von Beispielen ist im Allgemeinen ein schlecht gestelltes Problem, da die Abbildung nur für die Beispiele definiert ist [Pog90, Elg04].

Eine Möglichkeit die Einbettung zuverlässig zu lernen ist die Berücksichtigung der gelernten Mannigfaltigkeit der Beispiele. Auf der eingebetteten Mannigfaltigkeit ist eine robuste Interpolation zwischen den Punkten im niedrigdimensionalen Raum, also zwischen den Koordinaten der Beispiele im niedrigdimensionalen Raum, möglich, so dass eine Abbildung von der d-dimensionalen eingebetteten Mannigfaltigkeit zum Raum der Kovarianzdeskriptoren zuverlässig gelernt werden kann. Im Rahmen dieser Arbeit wird dieser Homöomorphismus entsprechend dieser Vorgehensweise durch RBF gelernt. Dazu wird das RBF-Interpolations-Rahmenwerk in [Elg08] verwendet, das durch die Arbeiten [Pog90, Bey96] motiviert wurde, worin RBF-Netzwerke zum Lernen nichtlinearer Abbildungen zwischen hochdimensionalen Räumen von Bildern und überwachten Parameterräumen verwendet werden. Das Lernen des Homöomorphismus auf Basis des RBF-Interpolations-Rahmenwerks gemäß [Elg08] wird im Folgenden vorgestellt.

Zum Lernen werden $e \times e$-Trainings-Kovarianzdeskriptoren $\mathcal{A} = \{\boldsymbol{\Sigma}_i\}$, $i = 1, \ldots, n$ durch $D = \frac{e \cdot (e+1)}{2}$-dimensionale Vektoren $\{\boldsymbol{s}_i\}$, $i = 1, \ldots, n$, $\boldsymbol{s}_i \in \mathbb{R}^D$ repräsentiert, wobei e der Dimension eines Merkmalsvektors entspricht und D somit der intrinsischen Dimension des Sym_n^+. Die Umrechnung erfolgt mittels dem folgenden $\mathrm{vec}_{\mathrm{OD}}$-Operator, der aus der oberen Dreiecksmatrix eines Kovarianzdeskriptoren einen D-dimensionalen Vektor bestimmt:

$$\boldsymbol{s} := \begin{pmatrix} \sigma_{1,1} \\ \sigma_{1,2} \\ \sigma_{2,2} \\ \sigma_{1,3} \\ \vdots \\ \sigma_{e,e} \end{pmatrix} = \mathrm{vec}_{\mathrm{OD}} \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \ldots & \sigma_{1,e} \\ \sigma_{2,1} & \sigma_{2,2} & \ldots & \sigma_{2,e} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{e,1} & \sigma_{e,2} & \ldots & \sigma_{e,e} \end{pmatrix} = \mathrm{vec}_{\mathrm{OD}}(\boldsymbol{\Sigma}) \,. \quad (5.10)$$

Die zugehörigen Vektoren im niedrigdimensionalen Raum sind durch $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\boldsymbol{y}_i \in \mathbb{R}^d$ gegeben. Die nichtlineare Abbildung, die im Folgenden als h_{RBF} bezeichnet ist, wird mittels RBF-Interpolanten $f^k : \mathbb{R}^d \to \mathbb{R}$ vom $\mathbb{R}^d$ zu allen Komponenten $1, \ldots, k, \ldots, D$ der zugehörigen D-dimensionalen Vektoren gelernt. Die RBF-Interpolanten sind durch

$$f_k(\boldsymbol{y}) = p_k(\boldsymbol{y}) + \sum_{i=1}^{n_u} r_{k,i} \phi\left(\mathrm{d}(\boldsymbol{y}, \boldsymbol{u}_i)\right) \quad (5.11)$$

definiert, wobei $\phi(\cdot)$ eine vorgegebene reellwertige Basisunktion ist. $r_{k,i}$ sind reelle Koeffizienten und $\boldsymbol{u}_i$ sind n_u frei wählbare Punkte auf der d-dimensionalen Mannigfaltigkeit ($n_u \leq n$), die nicht unbedingt Punkte aus $\boldsymbol{Y}$ sein müssen. $p_k(\boldsymbol{y})$ ist eine Linearkombination mit Koeffizienten $\boldsymbol{c}_k \in \mathbb{R}^{d+1}$:

$$p_k(\boldsymbol{y}) = [1 \quad \boldsymbol{y}^T] \cdot \boldsymbol{c}_k \,. \quad (5.12)$$

Als reellwertige Basisfunktion wird die gaußsche Funktion verwendet, die durch die Gleichung

$$\phi(s) = \mathrm{e}^{-\beta s^2}, \quad \beta > 0 \quad (5.13)$$

gegeben ist, wobei $\beta = 1$ gewählt wurde. Die Gesamtabbildung h_{RBF} von $\mathbb{R}^d \to \mathbb{R}^D$ berechnet sich dann aus den in Gleichung (5.11) definierten Interpolanten $f_k(\boldsymbol{y})$ durch

$$h_{\mathrm{RBF}}(\boldsymbol{y}) = \boldsymbol{B} \cdot \psi(\boldsymbol{y}) \,, \quad (5.14)$$

wobei $\boldsymbol{B}$ eine $D \times (n_u + d + 1)$ Matrix mit den Koeffizienten der D radialen Basisfunktionen ist:

$$\boldsymbol{B} = \begin{pmatrix} r_{1,1} & \dots & r_{1,n_u} & c_{1,0} & \dots & c_{1,d} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ r_{D,1} & \dots & r_{D,n_u} & c_{D,0} & \dots & c_{D,d} \end{pmatrix} . \tag{5.15}$$

Der Vektor $\psi(\boldsymbol{y})$ ist definiert durch

$$\psi(\boldsymbol{y}) := \left(\phi\left(\mathrm{d}(\boldsymbol{y}, \boldsymbol{u}_1)\right), \dots, \phi\left(\mathrm{d}(\boldsymbol{y}, \boldsymbol{u}_{n_u})\right), 1, \boldsymbol{y}^T\right)^T . \tag{5.16}$$

Um die Orthogonalität zu gewährleisten, wodurch das Problem gut konditioniert wird, wird die folgende Nebenbedingung eingeführt:

$$\sum_{i=1}^{n_u} r_{k,i}\, u_{j,i} = 0, \quad j = 0, \dots, d\,, \tag{5.17}$$

wobei $u_{j,i}$ die j-te Komponente von u_i für $i > 0$ ist und $u_{0,i} = 1$ gesetzt wird.

Die Matrix $\boldsymbol{B}$ ergibt sich aus der Lösung des linearen Gleichungssystems

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0}_{(d+1)\times(d+1)} \end{pmatrix} \boldsymbol{B}^T = \begin{pmatrix} \boldsymbol{S}^T \\ \boldsymbol{0}_{(d+1)\times D} \end{pmatrix} , \tag{5.18}$$

wobei $\boldsymbol{A} = (a_{i,j}) = \phi(\mathrm{d}(\boldsymbol{u}_i, \boldsymbol{u}_j))$, $i, j = 1, \dots, n_u$, und $\boldsymbol{P}$ eine $n_u \times (d+1)$ Matrix ist:

$$\boldsymbol{P} = \begin{pmatrix} 1 & \boldsymbol{u}_1^T \\ \vdots & \vdots \\ 1 & \boldsymbol{u}_{n_u}^T \end{pmatrix} \tag{5.19}$$

$\boldsymbol{S}$ ist eine $D \times n_u$ Matrix mit D-dimensionalen Repräsentanten (Vektoren) der Trainings-Kovarianzdeskriptoren $\mathcal{A}_u = \{\boldsymbol{\Sigma}_i\}$, $i = 1, \dots, n_u$:

$$\boldsymbol{S} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_{n_u})\,. \tag{5.20}$$

Die gelernte Abbildung repräsentiert somit einen Kovarianzdeskriptor durch eine Linearkombination nichtlinearer Basisfunktionen, die entlang der eingebetteten d-dimensionalen Mannigfaltigkeit an den Punkten $\boldsymbol{u}_i$, $i = 1, \dots, n_u$ zentriert sind.

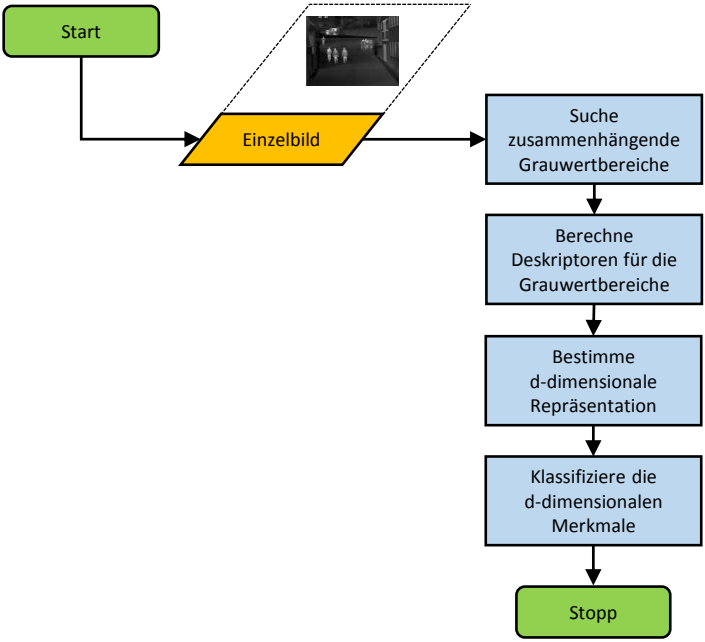


Abbildung 5.9: Übersichtsdiagramm über die Onlinephase.

Klassifikationsschritt

Die Menge der Kovarianzdeskriptoren für die ermittelten Bildregionen, die möglicherweise ein Körperteil zeigen, sei durch $\{\boldsymbol{t}_i\},\ i = 1, \ldots, n, \quad \boldsymbol{t}_i \in \mathbb{R}^D$ gegeben. Diese Deskriptoren werden dann klassifiziert, also einer Körperteilmannigfaltigkeit oder der Mannigfaltigkeit negativer Beispiele zugeordnet. Dazu werden die zur Laufzeit berechneten Kovarianzdeskriptoren mittels der nichtlinearen Abbildung aus dem zweiten Trainingsschritt auf die im ersten Trainingsschritt gelernte d-dimensionale Mannigfaltigkeit projiziert. Die niedrigdimensionalen Koordinaten der zu klassifizierenden Kovarianzdeskriptoren werden dabei mittels Umkehrung der gelernten Abbildung durch Lösen des folgenden Minimierungsproblems bestimmt (siehe [Elg08]):

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \ \mathrm{d}\left(\boldsymbol{t}_i, \boldsymbol{B} \cdot \psi(\boldsymbol{y})\right) . \tag{5.21}$$

Anschließend werden die niedrigdimensionalen Punkte mittels eines k-Nächste-Nachbarn-Klassifikators klassifiziert. Abbildung 5.9 zeigt eine Übersicht über die Onlinephase.

5.3 Verfahrensevaluation

Für den Nachweis, ob eine Verbesserung mit dem MaL basierten Ansatz erzielt werden kann, wurden zwei Evaluationen auf Bildausschnitten eines Wärmebilddatensatzes durchgeführt. Der Datensatz umfasst drei Körperteilklassen (Kopf-Schulter, Torso und Unterkörper) sowie negative Proben. Die für das Training des Klassifikators verwendeten Bildausschnitte der Körperteile und negativen Proben wurden mit Hilfe der Schwellwertfilterung extrahiert und manuell annotiert (Beispiele sind in der Abbildung 5.5 dargestellt). Für das Training wurden je Klasse 100 bzw. 250 Bildausschnitte betrachtet und als Testdaten standen je Klasse 50 bzw. 450 Bildausschnitte zur Verfügung. Die Originalbilder haben eine Auflösung von 640 Pixel (Breite) × 480 Pixel (Höhe). Die durchschnittliche Größe der Personen ist ca. 15 Pixel (Breite) × 65 Pixel (Höhe).

Abbildung 5.10: Beispielergebnisse der MaL basierten Körperteildetektion: Kopf- (grün), Torso- (rot) und Bein-Detektionen (gelb).

Evaluation 1. In [Met10] wurde für einen prinzipiellen Nachweis, ob eine Verbesserung mit dem MaL basierten Ansatz erzielt werden kann, eine Evaluation auf einem Teil des Wärmebilddatensatzes durchgeführt. Dazu wurde der MaL basierte Ansatz einmal mit der nichtlinearen Einbettung (Personenrepräsentation im niedrigdimensionalen Raum) und einmal ohne Einbettung (Personenrepräsentation durch Kovarianzdeskriptoren) durchgeführt. Bei dieser Evaluation wurden drei Körperteilklassen betrachtet: Kopf-Schulter, Torso und Unterkörper.

Für die Trainingsphase standen 400 annotierte Bildausschnitte zur Verfügung, jeweils 100 Ausschnitte pro Körperteilklasse und 100 negative Proben. Die Trainingsbildausschnitte wurden auf eine einheitliche Größe skaliert, sind jedoch — wie auch bei dem HOG-Ansatz — nicht einheitlich rotiert. Zur Bestimmung der d-dimensionalen Gesamtrepräsentation wurde $k = 25$ gewählt und für die nichtlineare Einbettung wurden die 20 nächsten Nachbarn betrachtet. Die Parameter resultieren aus Parameteroptimierungen und den Ergebnissen der theoretischen Untersuchungen von überwachten MaL Methoden für die Klassifikation in [Vur15]. Als Parameter d für die Gesamtrepräsentation der Körperteile wurde 12 gewählt.

Es wurde ein ausgewogener Testdatensatz betrachtet, der 200 Bildausschnitte umfasst: je Körperteil 50 und 50 zufällig ausgestanzte Bildregionen, die kein Körperteil zeigen und nicht durch die Schwellwertfilterung herausgefiltert wurden. Die Tabellen 5.1 und 5.2 zeigen Verwechslungsmatrizen,

		Grundwahrheit			
		Kopf-Schulter	Torso	Unterkörper	negative Klasse
Hypothese	Kopf-Schulter	50	0	0	3
	Torso	0	50	6	0
	Unterkörper	0	0	44	4
	negative Klasse	0	0	0	43

Tabelle 5.1: Verwechslungsmatrix: Ergebnis der kNN-Klassifikation der Kovarianzdeskriptoren ohne MaL.

		Grundwahrheit			
		Kopf-Schulter	Torso	Unterkörper	negative Klasse
Hypothese	Kopf-Schulter	50	0	0	2
	Torso	0	50	1	0
	Unterkörper	0	0	49	4
	negative Klasse	0	0	0	44

Tabelle 5.2: Verwechslungsmatrix: Ergebnis der MaL basierten kNN-Klassifikation.

die die Evaluationsergebnisse zusammenfassen: Tabelle 5.1 zeigt das Ergebnis der Klassifikation der Kovarianzdeskriptoren mittels kNN ohne MaL und Tabelle 5.2 zeigt das Ergebnis mit der MaL basierten Merkmalsextraktion. Durch MaL konnte ein besseres Klassifikationsergebnis erzielt werden. Abbildung 5.10 zeigt zwei qualitative Beispiele der MaL basierten Ergebnisse. Um die Ergebnisse zu bestätigen, wurde der Ansatz in einer zweiten Evaluation auf einem größeren Datensatz evaluiert und zusätzlich mit dem Basisverfahrens verglichen.

Evaluation 2. In der zweiten Evaluation werden die Ergebnisse des HOG-, Kovarianzdeskriptor- und MaL-Ansatzes gegenübergestellt. Es wurden wieder die Körperteilklassen Kopf-Schulter, Torso und Unterkörper für das Lernen der niedrigdimensionalen Gesamtrepräsentation betrachtet.

Die Trainingsdaten der 1000 annotierten Bildausschnitte: jeweils 250 positive Proben pro Körperteilklasse und 250 negative Proben (zufällige Bildausschnitte). Die Bildausschnitte wurden wieder auf eine einheitliche Größe skaliert und waren ebenfalls nicht einheitlich rotiert. Zur Bestimmung der d-dimensionalen Gesamtrepräsentation wurde aufgrund der größeren Trainigsdatenmenge $k = 80$ gewählt und bei der nichtlinearen Einbettung wurden die 5 nächsten Nachbarn betrachtet. Als Parameter d für die Gesamtrepräsentation der Körperteile wurde wieder 12 gewählt.

Es wurden wieder sowohl die Personenrepräsentation im niedrigdimensionalen Raum als auch die Personenrepräsentation durch Kovarianzdeskriptoren evaluiert. Zudem wurde zum Vergleich das HOG-Basisverfahren auf diesem Datensatz evaluiert. Zum Testen standen jeweils 450 Bildausschnitte je Körperteil sowie 450 durch die Schwellwertfilterung ausgestanzten Bildregionen, die kein Körperteil zeigen, zur Verfügung. Die Ergebnisse sind aus Sicht von 2-Klassenproblemen in der Tabelle 5.3 zusammengefasst.

Die auf Kovarianzdeskriptoren basierten Ansätze erzielten wie erwartet bessere Ergebnisse als das HOG-Basisverfahren. Außerdem bestätigt die zweite Evaluation das erste Ergebnis. Der niedrigdimensionale Ansatz (MaL basierte Repräsentation) erzielte in beiden Evaluationen — bis auf das Ergebnis der Torso-Klassifikation in der zweiten Evaluation — leicht bessere Ergebnisse. Die Ursache des schlechten Ergebnisses der Torso-Klassifikation lässt sich vermutlich auf die Konturen der Torsos zurückführen, die im

			Grundwahrheit	
			positiv	negativ
HOG	Hypothese	Kopf-Schulter	383	67
		Torso	274	176
		Unter-körper	377	73
		negative Klasse	29	421
Kovarianzdeskriptoren		Kopf-Schulter	443	7
		Torso	395	55
		Unter-körper	420	30
		negative Klasse	8	442
MaL basierte Repräsentation		Kopf-Schulter	448	2
		Torso	374	76
		Unter-Körper	427	23
		negative Klasse	6	444

Tabelle 5.3: Ergebnisse der jeweiligen Körperteil-Klassifikationen (kNN-Klassifikation) für die drei Repräsentationsansätze: HOG, Kovarianzdeskriptor und MaL basierte Repräsentation.

Vergleich zu den anderen Körperteilen weniger signifikante Eigenschaften haben. Eine signifikante Verbesserung resultiert nicht aus den Evaluationen, was teilweise auch an den guten Ergebnissen des Ansatzes mit den Kovarianzdeskriptoren liegt. Dennoch bekräftigen die Ergebnisse, dass die Inter-Diskriminanz zwischen unterschiedlichen Kovarianzdeskriptormannigfaltigkeiten durch die überwachte Strategie erhöht werden kann, das im Rahmen der erscheinungsbasierten Personenwiedererkennung weiter verfolgt wird (vgl. Kapitel 7).

5.4 Zusammenfassung

Ein wesentlicher Punkt, der durch den MaL-Ansatz bekräftigt wird, ist, dass die Annahme, dass Kovarianzdeskriptoren auf einer Mannigfaltigkeit liegen, vorteilhaft eingesetzt werden kann. Der Schwerpunkt der Untersuchungen lag auf kleinen Trainingsdatenmengen und niedrigdimensionalen Eingangsdaten (Kovarianzdeskriptoren mit kleinen Rängen), weshalb der Ansatz auch für *generische* Kovarianzdeskriptoren, die aus wenigen Merkmalen aufgebaut sind, anwendbar ist. Die Betrachtung unterschiedlicher Körperteilklassen lassen zudem erwarten, dass der Ansatz prinzipiell unabhängig von der Anwendungsdomäne ist, weshalb er auch in der erscheinungsbasierten Personenwiedererkennung wieder aufgegriffen wird (vgl. Kapitel 7). Die Annahme, dass die Kovarianzdeskriptoren auf einer Mannigfaltigkeit liegen, begegnet aufgrund der kleineren Intra-Diskriminanz zwischen den einzelnen Körperteilrepräsentationen zudem besser Herausforderungen, wie z.B unterschiedliche Kameraperspektiven oder ähnliche Erscheinungen. Damit können auch bei Verwendung von z.B. nicht-rotationsinvarianten Kovarianzdeskriptoren rotierte Körperteile robuster detektiert werden, zu denen es keine Trainingsdaten gibt.

6

Personentracking in Menschenmengen

In diesem Kapitel werden die theoretischen Ansätze des in Kapitel 4 vorgestellten mathematischen Rahmenwerks in ein videobasiertes Verfahren zum Tracking einzelner Personen in Menschenmengen umgesetzt [Met09]. Das Verfahren soll auf Luftbildaufnahmen angewandt werden, was bedeutet, dass überwiegend niedrig aufgelöste Bilder verarbeitet werden müssen. Für die Repräsentation der Personen werden Kovarianzdeskriptoren verwendet. Für die Assoziation, die Zuordnung der Personen zwischen den Bildern, wird die für die riemannsche Mannigfaltigkeit der Kovarianzdeskriptoren definierte Mahalanobis-Distanz verwendet. Außerdem fließen die in Kapitel 4 definierten statistischen Eigenschaften in den Assoziationsschritt mit ein.

Das Verfahren soll bei *Crowd and Riot Control* (CRC)-Einsätzen zur Eindämmung von Krawallen bei Menschenansammlungen eingesetzt werden. Unfriedliches Verhalten in Menschenansammlungen geht oft von einzelnen Personen aus, wie beispielsweise Rädelsführer, die andere Personen in der Menge aufhetzen. Eine Möglichkeit unfriedliche Situationen zu deeskalieren ist das Identifizieren und Entfernen solcher Störer aus der Menschenmenge. Dabei spielt die videogestützte Überwachung der Menge eine große Rolle.

Durch die videogestützte Überwachung der Menge können die Zielpersonen ausgemacht und geeignete Eingriffspunkte bestimmt werden. Das Ziel dieser Arbeit ist ein videobasiertes Verfahren, das zuverlässig einzelne manuell markierte Personen in Menschenmengen verfolgen kann. Im Rahmen der Evaluation wird die Zuverlässigkeit des Verfahrens anhand eines Datensatzes mit Luftbildaufnahmen von CRC-Einsätzen evaluiert. In Abbildung 6.1 sind beispielhaft acht Bilder aus diesem CRC-Datensatz dargestellt.

Das Verfahren ist in erster Linie für Luftbildaufnahmen erarbeitet worden, in denen mehrere Individuen aus einer Nadir-Sicht getrackt werden sollen. Eine große Herausforderung beim Tracking einzelner Personen in Luftbildern mit Menschenmengen ist in der Regel die große Distanz der videosensortragenden Plattform zur Menschenmenge. Die Durchmesser der Köpfe sind meistens kleiner als 10 Pixel, so dass oft nur kleine Körperflächen aus der Nadir-Sicht sichtbar sind. Zudem ist aus diesem Sichtwinkel die Diskriminanz der Körperflächen von unterschiedlichen Personen sehr gering, da sich die Draufsicht verschiedener Personen in der Regel nicht so stark unterscheidet wie deren Seitenansicht (siehe beispielsweise die blau eingekreisten und mit der Ziffer 1 gekennzeichneten Personen in Abbildung 6.2). Dies führt oft zu einer hohen Verwechslungsgefahr von Personen in der Menschenmenge. Durch Einschränkung des Suchbereichs könnte man der Verwechslungsgefahr zwar entgegenwirken, ob und inwieweit sich ein Suchbereich allerdings einschränken lässt, hängt u.a. von der Bildwiederholrate, Abstand der Sensorplattform zur Menschenmenge und der Fortbewegungsgeschwindigkeit der Person ab. Ein kleiner Suchbereich, der gerade so groß ist, dass immer nur der Kopf der im Video zu verfolgenden Person enthalten ist, wäre wünschenswert. In diesem Fall müsste die Person nicht mit den Personen in ihrer Umgebung verglichen werden, zumindest solange der Kopf nicht durch Köpfe anderer Personen verdeckt werden würde. Die Nadir-Sicht ist hinsichtlich Verdeckungen zwar sehr vorteilhaft, da es selten zu Verdeckungen durch andere Personen kommt, aufgrund variabler Parameter wie Fortbewegungsgeschwindigkeit der Menschenmenge etc., die in der Regel nur unzureichend prädiziert werden können, ist eine starke Einschränkung des Suchbereichs im Allgemeinen jedoch nicht möglich. Das Risiko, die Person aus dem Suchbereich zu verlieren, wäre zu groß. Aus diesem Grund ist die Berücksichtigung einer hohen Diskriminanz bei der Personenrepräsentation wesentlich bei der Auswahl der Deskriptoren.

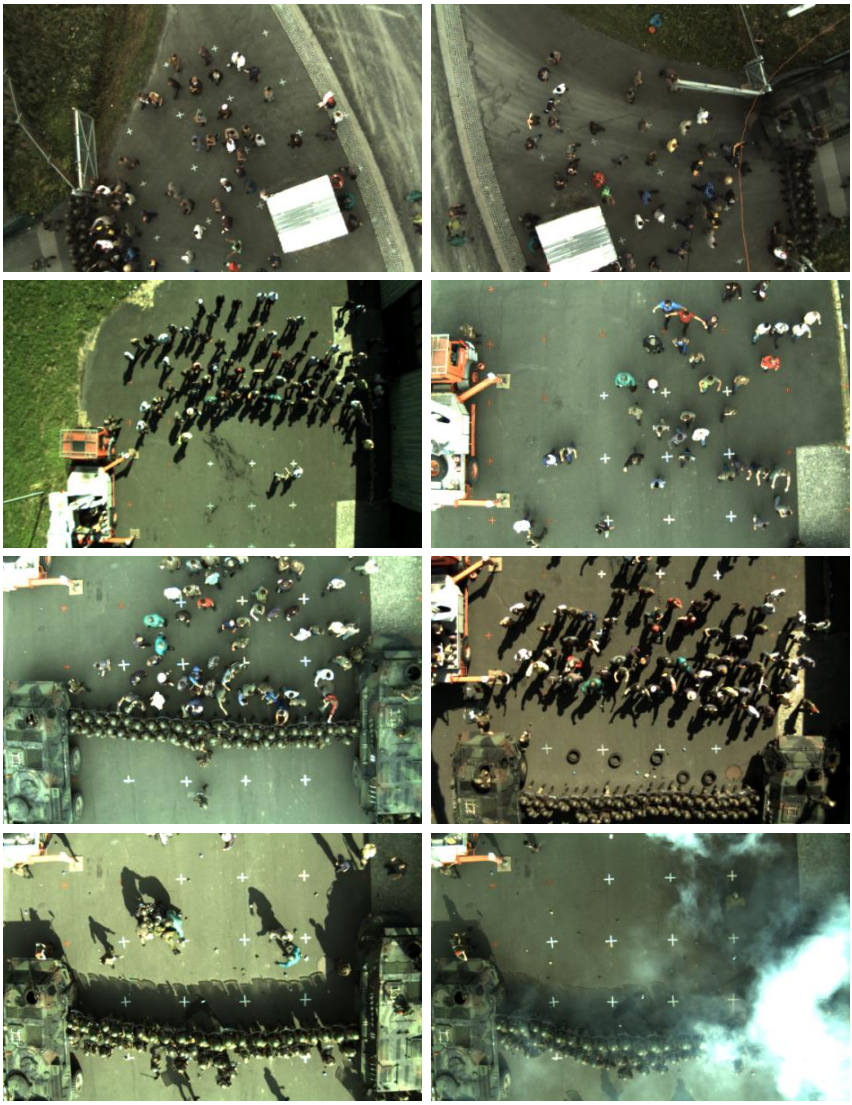


Abbildung 6.1: Beispielbilder aus dem CRC-Datensatz.

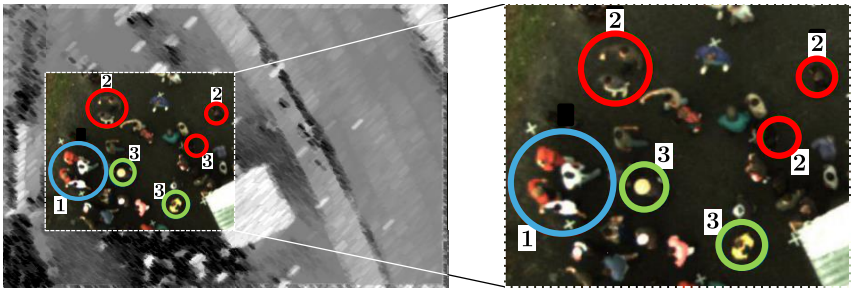


Abbildung 6.2: Beispiele für unterschiedliche Herausforderungen bei dem hier verwendeten CRC-Datensatz. Der markierte Bildausschnitt im linken Bild ist rechts vergrößert dargestellt. Bei den blau eingekreisten und mit der Ziffer 1 gekennzeichneten Personen besteht aufgrund der ähnlichen Erscheinungen zwischen jeweils zwei Personen eine erhöhte Verwechslungsgefahr. Diese Situationen sind auch in realen CRC-Szenarien aufgrund einheitlicher Kleidung, welche beispielsweise die Zugehörigkeit zur Gruppe zeigen sollen, oft anzutreffen. Eine andere Herausforderung sind die rot eingekreisten und mit der Ziffer 2 gekennzeichneten Personen, weil sie einen schwachen Kontrast zum Untergrund haben. Trackingverfahren verlieren solche Personen in vielen Fällen und bleiben meistens auf dem Hintergrund hängen. Beispiele für Personen, die hingegen einfach zu tracken sind, sind durch die grünen Kreise (Ziffer 3) gekennzeichnet.

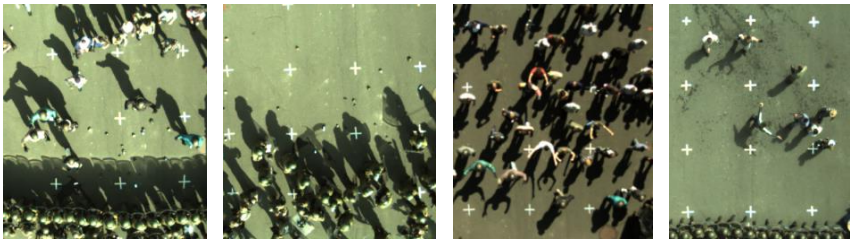


Abbildung 6.3: Beispiele für ungünstige Beleuchtungssituationen. Insbesondere Schatten stellen aufgrund ihrer Größe in der Nadir-Sicht oft eine große Herausforderung dar.

Verdeckungen von Personen durch andere Personen treten in der Nadir-Sicht in der Regel selten auf, was die Aufgabe des Personentrackings zwar erleichtert, allerdings können Personen durch Bäume, Laternen, Dächer etc. verdeckt werden. Diese Art von Verdeckung kann aber einfacher erkannt werden als eine Verdeckung durch eine oder mehrere Personen, da sich die Objekte in den meisten Fällen deutlich von Personen unterscheiden. Trackingverfahren, die fortlaufend ihr Objektmodell, im vorliegenden Fall die kovarianzdeskriptorbasierte Personenrepräsentation, aktualisieren, sollten Verdeckungen automatisch erkennen können, damit das zu trackende Modell nicht durch falsche Beobachtungen aktualisiert wird.

Eine weitere Herausforderung ist das Tracking von Personen, die einen schwachen Kontrast zum Untergrund haben. Beispiele hierfür sind die Personen, die durch die roten Kreise und mit der Ziffer 2 in der Abbildung 6.2 gekennzeichnet sind. Da sie kaum vom Untergrund zu unterscheiden sind, steigt die Gefahr, dass Trackingverfahren die Personen verlieren und auf dem Untergrund hängen bleiben. Personen, die einen hohen Kontrast zum Untergrund haben, wie beispielsweise die grün eingekreisten und mit der Ziffer 3 gekennzeichneten Personen in Abbildung 6.2, sind hingegen sehr zuverlässig zu tracken. Bei diesen Personen kommt außerdem erleichternd hinzu, dass sich ihre Kontur bzw. Jackenfarbe von der restlichen Menge unterscheidet. Diese Klasse von Personen, die einen hohen Kontrast zum Untergrund haben und sich von der restlichen Menge unterscheiden, stellt die einfachste Aufgabe für Trackingverfahren dar.

Eine weitere Schwierigkeit beim Tracking einzelner Personen in Luftbildern stellen bestimmte Beleuchtungssituationen und schnelle Beleuchtungsänderungen dar. Insbesondere Schatten können eine große Herausforderung darstellen, da sie aus der Nadir-Sicht oft deutlich größer als die Körperfläche der Person selbst erscheinen. Beinhaltet die Personenrepräsentation den Schatten oder auch nur einen Teil davon, steigt die Verwechslungsgefahr an, da sich die Schatten unterschiedlicher Personen nur gering anhand ihrer Kontur unterscheiden (siehe Abbildung 6.3). Außerdem sind schnelle Beleuchtungsänderungen problematisch für Trackingverfahren, da sie die Erscheinung von Personen schlagartig ändern können. Beleuchtungsänderungen können beispielsweise Farbwerte so drastisch ändern, dass die zu trackende Person von einem Bild zum nächsten nicht mehr anhand der Farbe korrekt zugeordnet werden kann.

Im Folgenden wird das Kovarianz-Trackingverfahren vorgestellt, das im Rahmen dieser Arbeit zum Tracking von Personen in Menschenmengen erarbeitet wurde. Es basiert auf dem Kovarianz-Tracker von Porikli et al. [Por06b] (siehe Abschnitt 6.1). Das in dieser Arbeit vorgestellte Verfahren ist als Single-Target-Tracking realisiert. Es können auch mehrere Personen gleichzeitig getrackt werden, was jedoch getrennt voneinander stattfindet. Der Tracker arbeitet auf Farbbildern (RGB) und verwendet Kovarianzdeskriptoren zur Repräsentation der zu trackenden Personen (Bildausschnitte). Die Deskriptoren basieren auf Farb- und Gradientenmerkmalen, die zusätzlich mit den Pixelkoordinaten der Bildausschnitte in Bezug gesetzt werden. Die Verwendung von Kovarianzdeskriptoren zum Tracking wurde erstmals 2006 in [Por06b] vorgeschlagen und über die Jahre in zahlreichen Arbeiten angepasst (vgl. Abschnitt 2.1.6). Wie in Kapitel 3 aufgeführt, sind Kovarianzdeskriptoren für die Repräsentation niedrig aufgelöster Personen geeignet, robust gegenüber Beleuchtungs- und Farbänderungen sowie diskriminant, weshalb diese Repräsentationsart einige der oben angesprochenen Herausforderungen begegnet. Zwei Herausforderungen die sich im CRC-Datensatz, der im Rahmen der Evaluation betrachtet wurde, stellen und gesondert gelöst werden müssen, sind Schatten sowie Personen, die einen schwachen Kontrast zum Untergrund haben. Insbesondere der letzte Fall stellt eine große Problematik dar, weshalb der Fokus bei der Erarbeitung der folgenden Adaptionen und Erweiterungen auf dieser Herausforderung lag.

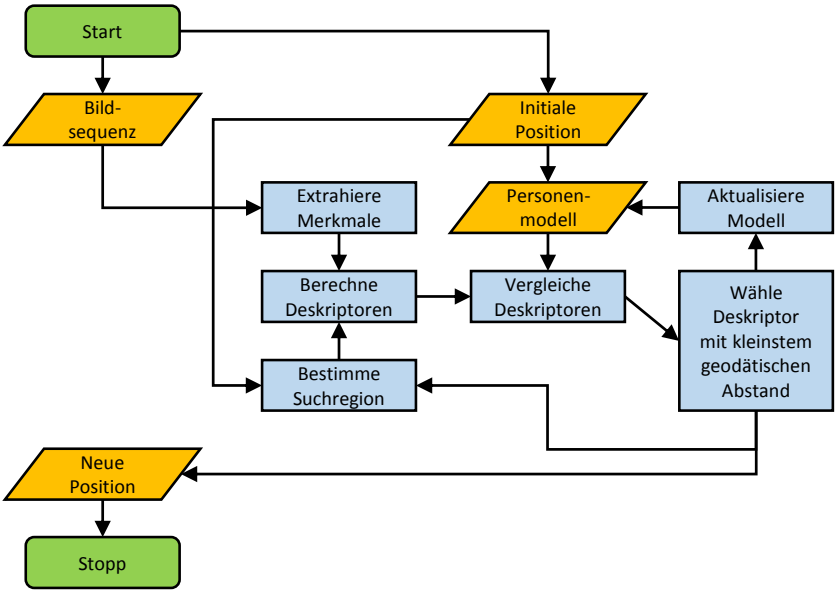


Abbildung 6.4: Flussdiagramm des Kovarianz-Trackingverfahrens von Porikli et al. [Por06b].

Zunächst wird im folgenden Abschnitt der Kovarianz-Tracker von Porikli et al. kurz dargestellt, bevor in Abschnitt 6.2 die Adaptionen und Erweiterungen des Verfahrens vorgestellt werden, die im Rahmen dieser Arbeit erarbeitet wurden. Der Basis-Kovarianz-Tracker von Porikli et al. wurde minimal abgeändert, weswegen dieser im Folgenden als *angepasster Porikli'scher Kovarianz-Tracker* (aPKov) bezeichnet wird.

6.1 Porikli'scher Kovarianz-Tracker (aPKov)

Der Porikli'sche Kovarianz-Tracker aPKov ist durch das Flussdiagramm in Abbildung 6.4 zusammenfassend dargestellt.

Im ersten Schritt (Initialisierungsphase) wird die zu trackende Person durch Setzen eines Rechtecks $\boldsymbol{R}$ im aktuellen Videobild $\boldsymbol{I}$ markiert. Dies kann manuell durch eine Person durchgeführt werden oder automatisch mittels eines Detektionsverfahrens geschehen. Der Mittelpunkt von $\boldsymbol{R}$ entspricht der initialen Position der Person im Bild. Nach dem Setzen des Rechtecks $\boldsymbol{R}$ wird ein Kovarianzdeskriptor für diesen Bildausschnitt berechnet, der die initiale Personenrepräsentation darstellt. Angelegt an Tuzel et al. [Por06b] wird in diesem Kapitel die Personenrepräsentation auch als Personenmodell bezeichnet, da sich die Personenrepräsentation während des Trackings aktualisieren kann.

Zur Berechnung des Kovarianzdeskriptors für die Koordinate (x, y) bzgl. der linken oberen Ecke des Bildausschnitts mit der Koordinate $(0, 0)$ wird der Merkmalsvektor $\boldsymbol{f}_{(x,y)}$ definiert, der sich aus x- und y-Pixelkoordinaten des Bildausschnitts sowie aus Farb- (RGB-Pixelwerten) und Gradientenmerkmalen bestimmt:

$$\boldsymbol{f}_{(x,y)} = \begin{pmatrix} X(x,y) \\ Y(x,y) \\ R(x,y) \\ G(x,y) \\ B(x,y) \\ \left|\frac{\delta \boldsymbol{I}_{\boldsymbol{R}}(x,y)}{\delta x}\right| \\ \left|\frac{\delta \boldsymbol{I}_{\boldsymbol{R}}(x,y)}{\delta y}\right| \\ \left|\frac{\delta^2 \boldsymbol{I}_{\boldsymbol{R}}(x,y)}{\delta x^2}\right| \\ \left|\frac{\delta^2 \boldsymbol{I}_{\boldsymbol{R}}(x,y)}{\delta y^2}\right| \end{pmatrix}, \tag{6.1}$$

mit $X(x, y) = x$ und $Y(x, y) = y$.

Sowohl in der Initialisierungsphase als auch während des Trackings kann zu jedem Zeitpunkt eine weitere Person zur Verfolgung markiert werden, die dann unabhängig von den anderen markierten Personen im Video getrackt wird.

Zur Laufzeit des Trackings (Onlinephase) wird in den darauffolgenden Videobildern nach der markierten Person gesucht. Dazu wird jeweils im folgenden Videobild ein Rechteck, das die gleiche Größe wie $\boldsymbol{R}$ hat, durch einen Suchbereich mit fester Größe um die aktuelle Bildposition geschoben. Die Größe entspricht dem initialen Rechteck. Nach jedem Verschiebungsschritt wird ein Kovarianzdeskriptor für den neuen Bildausschnitt berechnet und mit dem Kovarianzdeskriptor, der das Personenmodell $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}$ repräsentiert, verglichen. Der Deskriptor mit dem kleinsten geodätischen Abstand zu $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}$ bestimmt die neue Position der markierten Person. Der geodätische Abstand ergibt sich wie folgt aus der Gleichung (4.8) in Abschnitt 4.1.1:

$$\mathrm{g}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \sqrt{\langle \log_{\boldsymbol{\Sigma}_1}(\boldsymbol{\Sigma}_2) \mid \log_{\boldsymbol{\Sigma}_1}(\boldsymbol{\Sigma}_2) \rangle_{\boldsymbol{\Sigma}_1}} \,. \tag{6.2}$$

Die neuen Positionskoordinaten der markierten Person im Bild entsprechen dem Mittelpunkt des zugehörigen Rechtecks des gefundenen Deskriptors.

Zusätzlich wird das Personenmodell $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}$ durch diesen Deskriptor aktualisiert. Das neue Personenmodell ergibt sich aus dem im aktuellen Bild gefundenen und den letzten m Kovarianzdeskriptoren (aus den letzten m Bildern). Der Parameter m kann dabei frei gewählt werden. Er bestimmt, wie schnell sich das Modell an eventuelle Änderungen in der Personenerscheinung anpasst. Je kleiner m gewählt wird, desto schneller passt sich das Modell an. Eine Anpassung sollte grundsätzlich durchgeführt werden, da es in der Regel ständig zu Änderungen kommt, z.B. aufgrund von Beleuchtungs- oder Blickwinkeländerungen etc. Das neue Personenmodell $\bar{\boldsymbol{\Sigma}}^{*}_{\boldsymbol{R}}$ ergibt sich aus der Gleichung (4.10) (Abschnitt 4.1.2):

$$\bar{\boldsymbol{\Sigma}}^{*}_{\boldsymbol{R}} = \exp_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}} \left(\frac{1}{m+1} \sum_{i=1}^{m+1} \log_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}}(\boldsymbol{\Sigma}_i) \right) . \tag{6.3}$$

Um die Gefahr einer schleichenden Anpassung der Personenrepräsentation der Zielperson auf eine andere Person zu mindern, wird im Vergleich zu dem Kovarianz-Trackingverfahren von Porikli et al. beim aPKov die Personenrepräsentation jedoch nur aktualisiert, falls der geodätische Abstand unter einem frei gewählten Schwellwert liegt. Andernfalls wird der Kovarianzdeskriptor verworfen und auch die Position der markierten Person nicht aktualisiert. Findet keine Aktualisierung statt, wird der Suchbereich im darauffolgenden Videobild vergrößert.

6.2 Erweiterter Kovarianz-Tracker (eKov)

In diesem Abschnitt werden die im Rahmen dieser Arbeit vorgenommenen Anpassungen und Erweiterungen des Kovarianz-Trackers von Porikli et al. vorgestellt, der im Folgenden als *erweiterter Kovarianz-Tracker* (eKov) bezeichnet wird. Wie in Abbildung 6.5 dargestellt, wurde der geodätische Abstand durch eine Mahalanobis-Distanz im Raum der Kovarianzdeskriptoren ersetzt. Außerdem ist nicht nur der Kovarianzdeskriptor mit dem kleinsten Abstand zum Personenmodell für die Positionsaktualisierung ausschlaggebend. Es werden auch die umgebenden Kovarianzdeskriptoren bei der Entscheidung, ob das Personenmodell und die Position aktualisiert werden sollen, berücksichtigt.

Analog zum oben vorgestellten Kovarianz-Tracker wird im ersten Schritt der Initialisierungsphase die zu trackende Person durch Setzen eines Rechtecks im aktuellen Videobild markiert. Anschließend wird in den darauffolgenden Bildern nach der markierten Person gesucht. Dazu wird ein kreisförmiges Fenster mit vordefiniertem Radius, der sich aus dem initialen Rechteck ergibt, durch einen Suchbereich geschoben, dessen Größe vorab festgelegt wird und optional durch einen lokalen optischen Fluss-Schätzer automatisch angepasst werden kann. Dabei wird die Größe des Suchbereichs von der Stärke des optischen Flusses bestimmt. Diese adaptive Anpassung liefert in der Regel eine deutlich kleinere Suchbereichsgröße, ohne die Person zu verlieren. Eine Invarianz gegenüber Skalierung kann durch Vergrößern und Verkleinern des Suchbereichs erreicht werden.

Eine vorgenommene Änderung an dem Kovarianz-Tracker ist der Austausch der Metrik für die Bestimmung der Ähnlichkeit von Kovarianzdeskriptoren (vgl. die Abbildungen 6.4 und 6.5): Das von Porikli et al. verwendete Abstandsmaß (Gleichung (6.2)) wird durch die in Gleichung (4.15) definierte Mahalanobis-Distanz ersetzt. Die Mahalanobis-Distanz ist für den Raum der Kovarianzdeskriptoren eindeutig definiert und wird zur Bestimmung des Abstands zwischen einer Beobachtung $\boldsymbol{\Sigma}$, die innerhalb des kreisförmigen Fensters liegt, und dem aktuellen Personenmodell $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}$ verwendet. Sie ergibt sich aus der Gleichung (4.15) (siehe Abschnitt 4.1.4):

$$\Omega_{(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}},\boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}})}(\boldsymbol{\Sigma}) = \sqrt{\mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}}(\overrightarrow{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}\boldsymbol{\Sigma}})^T \, \boldsymbol{Cov}^{-1}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}} \, \mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}}(\overrightarrow{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}\boldsymbol{\Sigma}})}\,. \qquad (6.4)$$

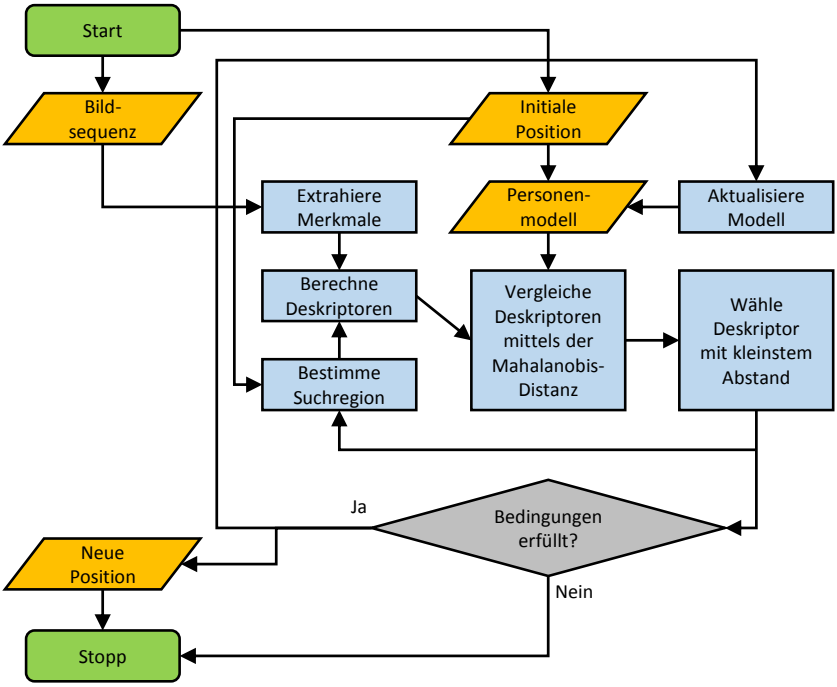


Abbildung 6.5: Flussdiagramm des eKov.

Der Operator $\mathrm{Vec}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}}$ ist der durch Gleichung (4.13) gegebene Isomorphismus. $\boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}}$ wird mittels der Gleichung (4.12) aus den Kovarianzdeskriptoren berechnet, die innerhalb des kreisförmigen Fensters liegen. Die Mahalanobis-Distanz wird für jede Position im Suchbereich berechnet. Die Position mit der kleinsten Mahalanobis-Distanz wird dann für eine Positionsaktualisierung in Betracht gezogen, wobei sowohl die Position als auch das Personenmodell $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}$ nur aktualisiert werden, wenn

- die Anzahl der Kovarianzdeskriptoren mit einer Mahalanobis-Distanz $< \delta$ größer als ein weiterer frei gewählter Schwellwert γ ist und
- die Varianz σ^2 der Normalverteilung $\mathcal{N}(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}, \boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{R}}})$ kleiner als eine fest vorgegebene Varianz ς ist.

Durch diese Bedingungen soll Ausreißern besser begegnet werden, was schließlich zu einer robusteren mittelwertbasierten Repräsentation führt.

6.3 Verfahrensevaluation

Als Grundlage für die Evaluation wurde eine CRC-Übung mit sechs Farbkameras aufgezeichnet, die auf einer 25 m hohen Kranplattform installiert und so ausgerichtet waren, dass sie die Personen aus der Nadir-Sicht aufgenommen haben. In Abbildung 6.6 sind Beispielbilder aus dem erstellten CRC-Datensatz zu sehen, die typische einsatzrelevante CRC-Szenarien zeigen, die in drei Eskalationsstufen eingeteilt werden können: Friedliche Demonstrationen, aufgebrachte Menschenmengen und Eskalationen, bei denen sich einzelne Personen gewaltsam verhalten und die Postenkette attackieren. Die akquirierten Bildsequenzen haben eine Bildwiederholrate von 20 Bildern pro Sekunde und die Auflösung eines Bildes ist 752×480 Pixel. Eine Person ist durchschnittlich 16×16 Pixel groß. Aufgrund der geringen Auflösungen der Personen und Verwendung von Farbkameras mit Bayer-Matrizen sind deutliche Farbsäume vorhanden — überwiegend an den Rändern der Personen (siehe Abbildung 1.4 rechts unten). Die Bildsequenzen zeigen Menschenmengen mit einer durchschnittlichen Größe von 60 Personen, worunter einige einen schwachen Kontrast zum Untergrund

(a) **(b)** **(c)**

Abbildung 6.6: Beispiele für die verschiedenen Eskalationsstufen: friedliche Demonstration **(a)**, aufgebrachte gestikulierende Menschenmenge **(b)** und eine Eskalationsstufe, bei der einzelne aggressive Personen die Postenkette attackieren **(c)**.

haben. Die Personen haben außerdem verschiedene Abstände zueinander und zeigen unterschiedliche Dynamiken.

Es wurden zwei quantitative Evaluationen des Kovarianz-Trackingverfahrens durchgeführt. Zunächst wurde der aPKov, wie er in Abschnitt 6.1 beschrieben ist, mit zwei anderen weit verbreiteten Trackingverfahren verglichen. Gegenstand der zweiten Evaluation war die um die statistischen Ergänzungen erweiterte Version des Porikli'sche Kovarianz-Trackers (eKov), der in Abschnitt 6.2 beschrieben ist. Der eKov wurde u.a. auf Basis der hier vorgestellten Evaluationsergebnisse des aPKov erarbeitet.

Zuerst werden die Ergebnisse des Vergleichs des aPKov mit zwei anderen Trackingverfahren ausführlich aufgeführt und gegenübergestellt. Neben dem aPKov wurde ein farbbasiertes Tracking von Merkmalen sowie ein Farbhistogramm-Tracker evaluiert. Beide Verfahren werden vor der Verfahrensevaluation kurz vorgestellt. Außerdem werden Metriken zur Evaluation der Trackingverfahren definiert.

Referenzverfahren 1: Farbbasierter Merkmals-Tracker. Der *farbbasierte Merkmals-Tracker* (FMT) ist ein um Farbe erweitertes Verfahren der pyramidalen Implementierung des *Kanade-Lucas-Tomasi* (KLT)-Merkmal-Trackers [Bou99]. Merkmals-Tracker wie der KLT-Tracker sind aufgrund

ihrer Robustheit heute noch sehr weit verbreitet [Rez17], oft auch in Kombination mit anderen Verfahren (siehe z.B. [Dao17]). Der *original* grauwertbasierte KLT-Tracker in [Bou99], der auf den Arbeiten in [Tom91, Shi94] basiert, extrahiert zunächst Ecken (Merkmale mit starken Gradienten in x- und y-Richtung) aus dem Bild (hier im Bildausschnitt der markierten Person). Anschließend wird jede Ecke im darauffolgenden Bild mittels Vergleich der umgebenden Bildstruktur gesucht. Dies wird dann von Bild zu Bild fortgesetzt, wobei in jedem Bild zusätzlich neue Ecken detektiert werden können. Die neue Position der Person wird relativ zur vorherigen Position durch Mittelung der optischen Flussvektoren, also aus den Geschwindigkeiten der Punkte, bestimmt. Im Rahmen dieser Arbeit wurde der KLT-Tracker von Kanade et al. folgendermaßen angepasst. Es werden nur Ecken auf der markierten Person getrackt, deren Farbe nicht oder nur selten in der Umgebung der Person vorhanden ist. Außerdem erfolgt die Mittelung der optischen Flussvektoren gewichtet. Je seltener die Farbe in der Umgebung der Person vorhanden ist, desto stärker geht der Flussvektor in die Berechnung ein.

Referenzverfahren 2: Farbhistogramm-Tracker. Farbhistogramme sind zum Tracking farbiger Bildregionen weit verbreitet (siehe z.B. [Yil06, Sal12, Xi13]). Sowohl die Berechnung als auch der Vergleich von Farbhistogrammen ist schnell und einfach zu implementieren. Der hier als Referenzverfahren verwendete *Farbhistogramm-Tracker* (FHT) verwendet normalisierte RGB-Farbhistogramme zur Repräsentation von Personen, die z.B. der Personenrepräsentation in [Woj02] ähnelt. Im Vergleich zu den konventionellen Ansätzen werden die einzelnen Farbklassen — ähnlich wie bei dem FMT — jedoch gewichtet. Farben, die auf der markierten Person oft und in deren Umgebung selten vorhanden sind, werden hoch gewichtet. Als Ähnlichkeitsmaß wird die Bhattacharyya-Distanz verwendet [Bha43].

6.3.1 Evaluations-Metriken

Bei den CRC Einsätzen ist es wichtig, dass eine in einem Bild markierte Person in darauffolgenden Bildern zuverlässig wiedererkannt wird, also

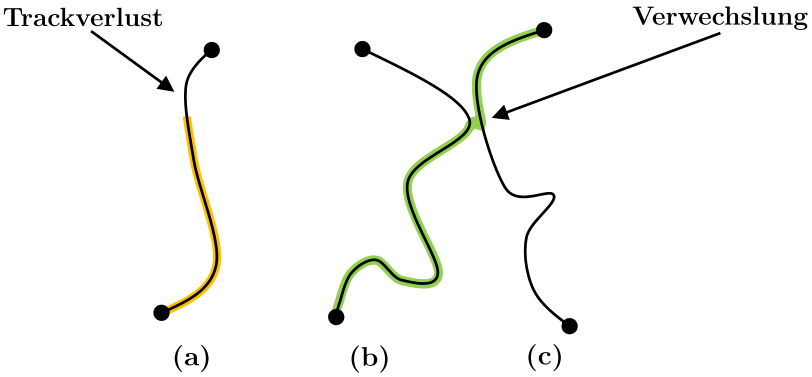


Abbildung 6.7: Trackverlust aufgrund von z.B. einer Verwechslung mit dem Hintergrund **(a)** und ein Überspringen des Tracks auf eine andere Person **(b)** → **(c)**.

dass es zu keinen Verwechslungen zwischen den Personen oder mit dem Hintergrund kommt, was — angelehnt an [Smi05] — den Tracking-Fehlern in Abbildung 6.7 entspricht. Außerdem ist eine präzise Positionsbestimmung der markierten Person in den Bildern wichtig. Aus diesen Zieldefinitionen ergeben sich folgende Forderungen an ein Verfahren zum Personentracking für CRC-Einsätze:

- die korrekte Wiedererkennung einer Person von Bild zu Bild und
- eine präzise Positionsbestimmung im Bild.

Metrik für die korrekte Wiedererkennung von Bild zu Bild. Die korrekte Wiedererkennung ist insbesondere beim Tracking einzelner Personen in Menschenmengen sehr wichtig und wird im Rahmen dieser Evaluationen mittels

$$\mathrm{RPR} = \frac{r_p}{r_p + f_n} \quad \text{und} \tag{6.5}$$

$$\mathrm{PV} = \frac{r_p}{r_p + f_p} \tag{6.6}$$

	Richtige Person	Andere annotierte Person oder Hintergrund
Tracker liefert Hypothese	r_p	f_p
Tracker liefert keine Hypothese	f_n	r_n

Tabelle 6.1: Basis-Metriken r_p (*richtig positiv*), f_p (*falsch positiv*), f_n (*falsch negativ*) und r_n (*richtig negativ*).

bestimmt (*Richtig-Positiv-Rate* (RPR) und *positiver Vorhersagewert* (PV)). Dabei sind r_p, f_n und f_p Basis-Metriken, deren Definitionen in der Tabelle 6.1 gegeben sind. Zur Vollständigkeit ist auch die Basis-Metrik r_n aufgeführt.

Die Basis-Metriken werden durch einen *Punkt-in-Polygon-Test* bestimmt. Dabei wird überprüft, ob die Hypothese eines Trackers (Punkt) innerhalb der Annotation (Grundwahrheit-Polygon) liegt, welche die gesuchte Person umrandet (vgl. Abbildung 6.8). Pro Person und Bild liefern die Trackingverfahren genau eine Hypothese. Die Polygone wurden mit dem in [Mül08] vorgestellten Annotationswerkzeug erstellt, wobei um jede Person von Hand ein Polygon gezeichnet und in die darauffolgende Bildern computergestützt fortgeführt wurde. Die Basis-Metriken werden für jedes Bild berechnet und auf die Gesamtzahl der Bilder normiert.

r_p gibt die Anzahl der korrekten Hypothesen an, also die Anzahl der Fälle, in denen die neu berechnete Position der Person innerhalb des korrekten Polygons liegt. f_p gibt die Anzahl der Fälle an, in denen die neue Position auf einer anderen Person (in einem anderen Polygon) oder auf dem Hintergrund liegt (Anzahl falscher Hypothesen). Die Anzahl der Fälle, in denen für die Zielperson keine Hypothese erstellt wurde, obwohl die zu trackende Person sichtbar ist, sind durch f_n gegeben und die Fälle, in denen keine Hypothesen gemacht werden und korrekterweise die Person auch nicht sichtbar ist, durch r_n. Abbildung 6.9 veranschaulicht anhand zweier Beispielbilder die vier unterschiedlichen Fälle.

Da die zu evaluierenden Trackingverfahren exakt eine Hypothese pro Bild liefern und in dem für die Evaluation verwendeten CRC-Datensatz in

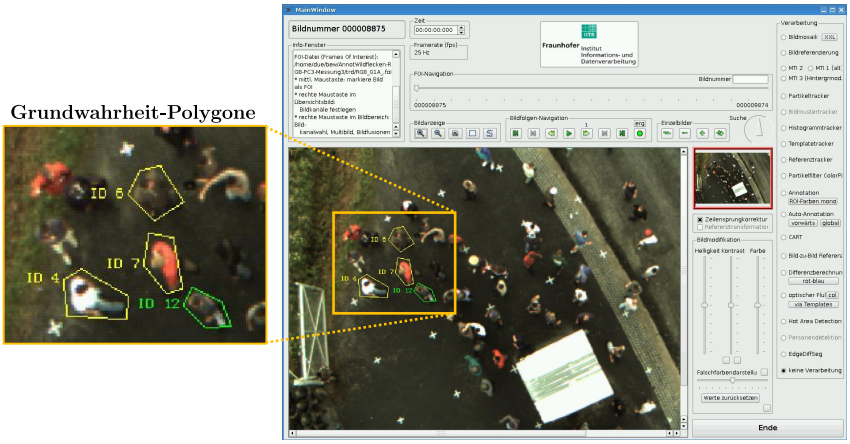


Abbildung 6.8: Bildschirmfoto des eingesetzten Annotationswerkzeugs mit einem vergrößerten Bildausschnitt, der beispielhaft vier annotierte Grundwahrheit-Polygone zeigt. In jedem Bild wurden alle Personen durch ein Polygon mit mindestens vier Ecken umrandet. Die Farben der Polygone sind im Rahmen dieser Evaluation irrelevant. Das grüne Polygon beispielsweise steht für das aktuell ausgewählte bzw. das zuletzt gezeichnete Polygon.

Abbildung 6.9: Übersicht über die vier Basis-Metriken. Ein Kreuz zeigt die berechnete Position und der Kreis gibt an, dass keine Position berechnet bzw. Positionsaktualisierung durchgeführt wurde.

jedem Bild alle Personen sichtbar und annotiert sind, entspricht in den folgenden beiden Evaluationen der Wert von f_p immer dem Wert von f_n. Wenn die Hypothese nicht auf der Person liegt entsteht gleichzeitig ein f_p, da die Hypothese nicht auf das Ziel zeigt, und ein f_n, da das Ziel nicht durch eine Hypothese abgedeckt ist. Die Basis-Metrik r_n ist für die folgenden Evaluationen nicht relevant, da alle Personen in allen Bildern sichtbar sind und r_n somit 0 ist.

Metrik für die Positionsunsicherheit. Die Unsicherheit bei der Positionsbestimmung wird anhand dem *Objekt-Tracking-Fehler* (OTF) berechnet. Der OTF ist eine Metrik für die Positionsunsicherheit von Trackingverfahren, die von Black et al. vorgeschlagen wurde [Bla03]. Die Unsicherheit bestimmt sich aus dem Abstand zwischen dem Schwerpunkt der Trackinghypothese — in unserem Fall die Punkt-Hypothese — mit den Koordinaten x_i^h und y_i^h und dem Schwerpunkt des Polygons, das die zu trackende Person umschließt, mit den Koordinaten x_i^s und y_i^s im i-ten Bild. Der OTF ist wie folgt definiert:

$$\mathrm{OTF} := \frac{1}{n_{hs}} \sum_{i=1}^{n_{hs}} \sqrt{\left(x_i^h - x_i^s\right)^2 + \left(y_i^h - y_i^s\right)^2}\,. \tag{6.7}$$

n_{hs} ist die Anzahl der Bilder, für die es sowohl die Hypothesen des Trackingverfahrens als auch die annotierten Personen gibt.

Neben dem mittleren OTF wird auch der $\mathrm{OTF}_{\min}$ (kleinste OTF) und $\mathrm{OTF}_{\max}$ (größte OTF) bei der Evaluation der Positionsunsicherheit bestimmt, die sich aus folgenden Gleichungen ergeben:

$$\mathrm{OTF}_{\min} = \min\left(\mathrm{OTF}_i\right), \tag{6.8}$$

$$\mathrm{OTF}_{\max} = \max\left(\mathrm{OTF}_i\right), \quad i = 1, \ldots, n_{hs}. \tag{6.9}$$

6.3.2 Evaluation des aPKov

In diesem Abschnitt werden die Ergebnisse des Vergleichs des aPKov mit den beiden konventionellen Trackingverfahren FMT und FHT vorgestellt [Hüb08]. Für die Evaluation standen 1001 aufeinanderfolgende Bilder

	FMT	**FHT**	**aPKov**
Anzahl der Grundwahrheiten je Person	1000	1000	1000
r_p	930,8	573,7	950,3
f_n	69,2	426,3	49,7
RPR	0,931	0,574	0,95

Tabelle 6.2: Mittlere RPR des Merkmals-Trackers (FMT), Farbhistogramm-Trackers (FHT) und aPKov.

zu Verfügung: Das *Bild 0* für die Initialisierung der Personen und 1000 darauffolgende mit jeweils 30 annotierten Personen für die Auswertung. Die Initialisierung erfolgte bei allen Trackingverfahren durch dasselbe Rechteck (fest vorgegebene Koordinaten). Die Tracker wurden nach einem Trackverlust nicht neu aufgesetzt. Die Personen sind jeweils durch ein Grundwahrheit-Polygon mit mindestens vier Ecken umrandet (ein Beispiel für ein Grundwahrheit-Polygon ist im linken Bild in Abbildung 6.9 gegeben). Die Auswertung erfolgte anhand der in Abschnitt 6.3.1 definierten Evaluations-Metriken. Im Folgenden werden die durchschnittlichen Evaluationsergebnisse, gemittelt über die 30 Personen, und zusätzlich die RPRn zu den einzelnen Personentracks aufgeführt.

Wiedererkennung von Bild zu Bild. Da $f_p = f_n$, entspricht der positive Vorhersagewert der RPR: $\frac{r_p}{r_p+f_p} = \frac{r_p}{r_p+f_n}$. Die mittleren RPRn der jeweiligen Trackingverfahren sind in der Tabelle 6.2 aufgeführt. Die mittlere RPR beim FHT ist deutlich niedriger als bei den anderen beiden Verfahren, die überwiegend nur bei den Personen mit einem schwachen Kontrast zum Hintergrund falsche Hypothesen lieferten (vgl. auch Abbildung 6.10).

Positionsunsicherheit. Für die Bestimmung der Positionsunsicherheit wurde der mittlere, kleinste und größte OTF ermittelt. Das jeweilige — über alle Personen gemittelte — Gesamtergebnis der einzelnen Trackingverfahren ist in der Abbildung 6.11 dargestellt.

Abbildung 6.10: Identifikationsnummern der 30 annotierten Personen (links). Die Bildausschnitte rechts zeigen beispielhaft vier Personen mit auffällig niedriger RPR.

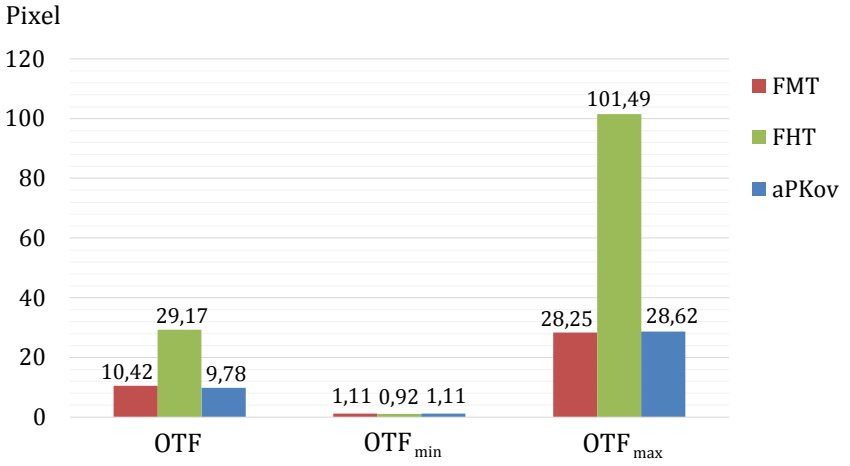


Abbildung 6.11: Mittlere, kleinste und größte Objekt-Tracking-Fehler der einzelnen Trackingverfahren. Die Objekt-Tracking-Fehler sind über alle Personen gemittelt.

	aPKov	**eKov**
Anzahl der Grundwahrheiten je Person	2000	2000
RPR	0,971	0,985
Objekt-Tracking-Fehler in Pixel	14,23	7,96

Tabelle 6.3: Mittlere RPRn und Objekt-Tracking-Fehler des aPKov und eKov.

Der FHT weist eine sehr großen OTF auf, der daraus resultiert, dass dieser Ansatz sehr oft die Zielperson verliert. Der FMT und aPKov weisen einen ähnlichen, niedrigen mittleren OTF auf.

6.3.3 Evaluation des eKov

In diesem Abschnitt werden die Ergebnisse des Vergleichs des aPKov mit dem eKov vorgestellt. Der Vergleich wurde für eine bessere Vergleichbarkeit ohne die automatische Suchraumanpassung durch den lokalen optischen Fluss-Schätzer durchgeführt und beide Verfahren wurden nach einem Trackverlust nicht neu aufgesetzt. Für den Vergleich der beiden Trackingvarianten standen neben dem Bild für die Initialisierung der Personen weitere 2000 annotierte aufeinanderfolgende Bilder zur Verfügung, in denen ebenfalls alle Personen durch ein Grundwahrheit-Polygon mit mindestens vier Ecken umrandet sind. Bei der Evaluation wurden alle Bilder bzw. Grundwahrheit-Polygone berücksichtigt, also auch dann, wenn der eKov keine Positionsaktualisierung durchgeführt hat. In diesem Fall wurde die letzte aktualisierte Position betrachtet.

In der Tabelle 6.3 sind die durchschnittlichen Evaluationsergebnisse aufgeführt. Sowohl bei der RPR als auch beim OTF konnte eine Verbesserung erreicht werden, wobei der OTF beim eKov um fast die Hälfte des aPKov-OTF gesenkt werden konnte.

Eine genauere Betrachtung der Metriken hinsichtlich einzelner Personen verdeutlicht die Problematik eines schwachen Kontrasts zwischen einer Person und dem Untergrund. Anhand der einzelnen Auswertungen wurden

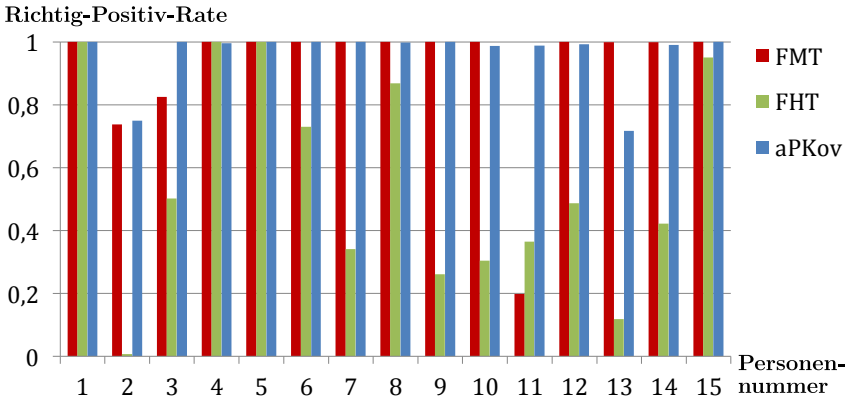


Abbildung 6.12: RPRn der einzelnen Personentracks 1 - 15.

	aPKov	**eKov**
Anzahl der Grund-wahrheiten je Person	2000	2000
RPR	0,558	0,749
Objekt-Tracking-Fehler in Pixel	76,31	14,11

Tabelle 6.4: Mittlere RPRn und Objekt-Tracking-Fehler des aPKov und eKov für die vier Personen, die rechts in der Abbildung 6.10 zu sehen sind.

die Personen mit auffällig niedriger RPR bestimmt und gesondert evaluiert. Die RPR der einzelnen Personentracks sind für alle drei Trackingverfahren in den Abbildungen 6.12 und 6.13 dargestellt.

Die Personen, die durch die Nummern $2, 11, 13$ und 22 gekennzeichnet sind, wurden aufgrund der niedrigen RPR zusätzlich gesondert evaluiert. Die über die vier Personen gemittelten Ergebnisse sind in der Tabelle 6.4 zusammengefasst. Das Ergebnis zeigt, dass auch bei problematischen Fällen, wie in diesem Fall die Personen mit einem schwachen Kontrast zum Hintergrund, ein deutlich besseres Trackingergebnis erwartet werden kann.

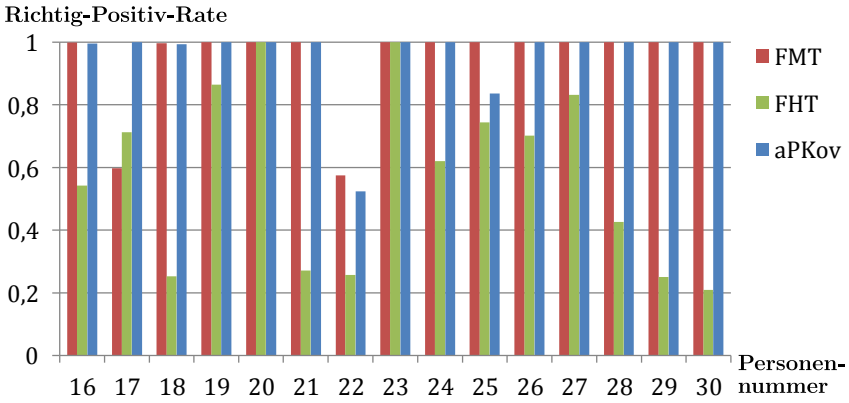


Abbildung 6.13: RPRn der einzelnen Personentracks 16 - 30.

6.4 Zusammenfassung

Im Vergleich zu bisherigen Kovarianzdeskriptor basierten Trackingverfahren wurde die im Sym_n^+ definierte Mahalanobis-Distanz im Assoziationsschritt verwendet und eine neue Aktualisierungsstrategie vorgeschlagen, die anhand statistischer Eigenschaften einzelne Kovarianzdeskriptoren in diesem Schritt ausschließt. Die Ergebnisse zeigen, dass dadurch den anwendungsbezogenen Herausforderungen besser begegnet werden kann. Es konnten insbesondere Verbesserungen hinsichtlich der Problematik des schwachen Kontrasts von Personen zum Hintergrund erzielt und die Verwechslungsgefahr von Personen, die eine ähnlichen Erscheinung haben, gesenkt werden, was u.a. schleichende Übergänge zwischen Personen besser vermeidet.

7

Erscheinungsbasierte Personenwiedererkennung

In diesem Kapitel werden die theoretischen Ansätze des in Kapitel 4 vorgestellten mathematischen Rahmenwerks und die Erkenntnisse aus den beiden vorherigen Kapiteln (Personendetektion und -tracking) in ein videobasiertes Verfahren zur automatischen Wiedererkennung von Personen umgesetzt [Met12a, Met12b, Met14]. Die Tracking-Ergebnisse in Kapitel 6 zeigen, dass eine Person robust durch einen Mittelwert repräsentiert werden kann, der sich aus mehreren Kovarianzdeskriptoren berechnet. Dabei ist es allerdings wichtig, dass keine *Ausreißer* mit einfließen, was durch die eKov-Ergebnisse verdeutlicht wird. Darüber hinaus zeigt das Kapitel 5, dass in Fällen, in denen Kovarianzdeskriptoren ähnliche Bildausschnitte repräsentieren, die Annahme getroffen werden kann, dass Kovarianzdeskriptoren eine Untermannigfaltigkeit im Sym_n^+ bilden. Diese Erkenntnis wird in einer einzelbildbasierten Neusortierung wieder aufgegriffen.

Das Ziel der Wiedererkennung ist es, eine Person, die manuell ausgewählt wurde, in Videodaten einer oder unterschiedlicher Bild- oder Videoquellen unabhängig von Aufnahmeort und -zeit wiederzufinden. Im Rahmen dieser Arbeit liegt der Fokus auf Videodaten, die von Kameranetzwerken oder aus polizeilich erfasstem Bildmaterial stammen. Mögliche Einsatzgebiete des hier vorgeschlagenen Wiedererkennungsverfahrens sind beispielsweise

Abbildung 7.1: Ausschnitte von Beispielsequenzen sechs unterschiedlicher Personen, für die das Wiedererkennungsproblem gelöst werden soll. Die erste Reihe zeigt Sequenzen aus dem Kameranetzwerk-Datensatz und die zweite Reihe aus dem Fahndungsdatensatz, die für die Evaluation des hier vorgestellten Verfahrens verwendet wurden. Die Bildausschnitte des Kameranetzwerk-Datensatzes sind im Vergleich zum Fahndungsdatensatz deutlich niedriger aufgelöst. Zudem wurde der Hintergrund bei dem Kameranetzwerk-Datensatz während des Trackings durch ein Bewegungsdifferenzverfahren entfernt [Met14].

die Unterstützung von Sicherheitskräften und Polizeien in sicherheitsrelevanten Infrastrukturen, wie z.B. Bahnhöfen oder Flughäfen, und die Beweissicherung, bei der Straftäter in Videodatenbanken gesucht werden.

Da die Anzahl der zu beobachtenden bzw. auszuwertenden Kameras seit Jahren rasant zunimmt, scheidet eine rein manuelle Wiedererkennung schon allein aufgrund unzureichender Personalkapazitäten aus. Deshalb wurde im Rahmen dieser Arbeit das Ziel verfolgt, ein hochautomatisiertes Wiedererkennungswerkzeug mit hoher Genauigkeit zu erarbeiten. Das Verfahren wurde für Videodaten, in denen die Personen niedrig aufgelöst sind, entworfen. Obwohl die Anschaffungskosten hochauflösender Kameras sinken und die Auflösungen laufend steigen, werden im Bereich der Videoüberwachung aus wirtschaftlichen Gründen auch zukünftig noch Videodaten mit gering aufgelösten Personen verarbeitet. In diesen Fällen können aufgrund der Auflösung oft keine biometrischen Ansätze eingesetzt

werden (vgl. Abschnitt 3.1). Das hier vorgestellte Verfahren sucht die Personen anhand ihrer Erscheinung in den Bilddaten, d.h. anhand ihrer Kleiderfarbe, Muster auf der Kleidung und Accessoires, etc.

Bei der Erarbeitung des Wiedererkennungsverfahrens lag der Fokus auf dem Vergleich einer oder mehrerer Bildsequenzen einer Person mit Bildsequenzen aus einem Kameranetzwerk oder einer vorhandenen Videodatenbank. Ein Vergleich mit Einzelbildern ist auch möglich, der im Rahmen dieser Arbeit zur Bestätigung und Korrektur des Vergleichsergebnisses des sequenzbasierten Ansatzes durchgeführt wird.

Das Verfahren wurde in erster Linie für Seiten- und Schrägansichten von Personen erforscht. Die Wiedererkennung von Personen in Nadir-Sichten ist auch möglich, allerdings sinkt dabei im Allgemeinen die Wahrscheinlichkeit einer korrekten Wiedererkennung aufgrund der von oben nicht sichtbaren Körperfläche. Vergleiche dazu in Kapitel 8 die Tabelle 8.6, die das Ergebnis einer einzelbildbasierten Evaluation auf einem Luft-Boden-Bilddatensatz zusammenfasst.

Die Abbildung 7.1 zeigt einige Beispiele für Sequenzen von Bildausschnitten einzelner Personen, auf die im Rahmen der erscheinungsbasierten Wiedererkennung der Fokus gelegt wurde. Die Sequenzen wurden automatisch durch ein Trackingverfahren erstellt [Met14], in dem automatisch detektierte bzw. getrackte Personen rechteckig in den Bildern markiert wurden.

Eine Übersicht über den hier vorgestellten erscheinungsbasierten Wiedererkennungsansatz ist in Abbildung 7.2 dargestellt. Als Eingabe werden Sequenzen achsparalleler Rechtecke um Detektionshypothesen verwendet. Eine vom Trackingverfahren erzeugte zusammenhängende Sequenz solcher Rechtecke wird im Folgenden als Tracklet bezeichnet, wobei zwischen zwei Arten von Tracklets unterschieden wird. Sequenzen, die aus einer Datenbank oder von einem Multi-Kamera-Netzwerk stammen, werden als Galerie-Tracklets und eine Sequenz, die mit den Galerie-Tracklets abgeglichen werden soll, wird als Anfrage-Tracklet bezeichnet.

Die einzelnen Verarbeitungsschritte des Verfahrens sind im Folgenden der Reihe nach zusammengefasst:

1. Kovarianzdeskriptoren für ausgestanzte Personenbildausschnitte (Einzelbilder) berechnen.

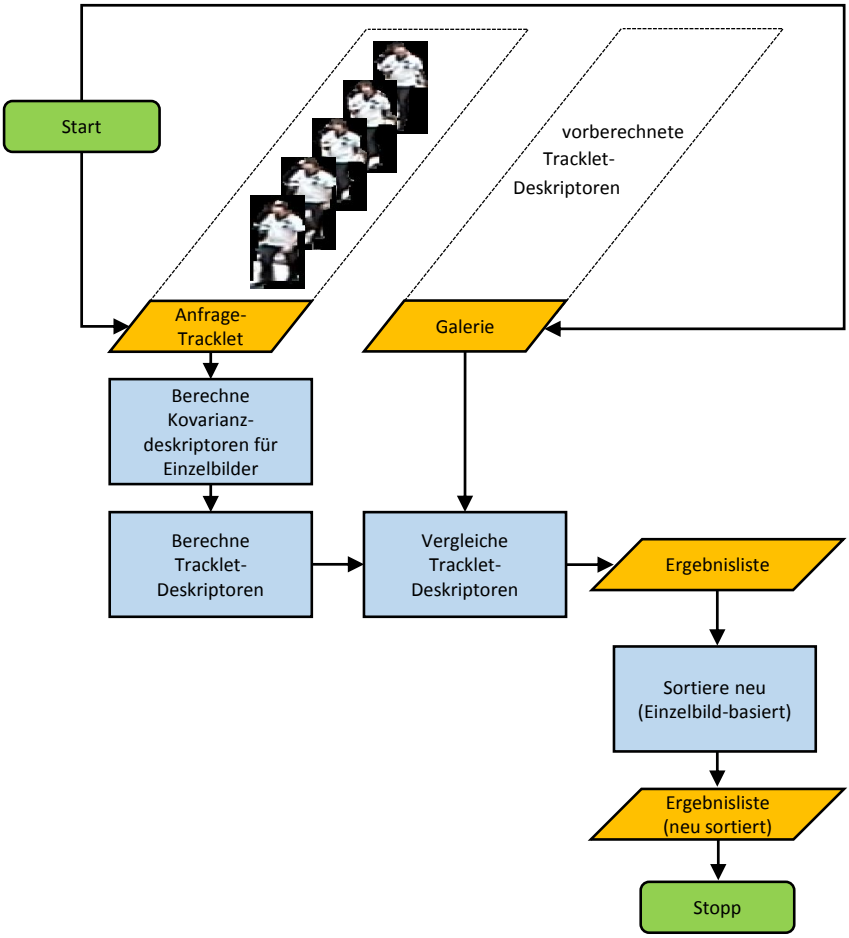


Abbildung 7.2: Flussdiagramm des erscheinungsbasierten Wiedererkennungsansatzes zur Laufzeit. Ein Anfrage-Tracklet wird mit Galerie-Tracklets anhand von Tracklet-Deskriptoren verglichen. Das Ergebnis ist eine Liste mit Galerie-Tracklets, die anhand der Ähnlichkeit zum Anfrage-Tracklet sortiert ist.

2. Trackletbasierte Personenrepräsentation aus den Kovarianzdeskriptoren berechnen.
3. Vergleich des Anfrage-Tracklets mit den Galerie-Tracklets anhand der Tracklet-Deskriptoren und Auflistung der Vergleichsergebnisse, sortiert nach ihrer Ähnlichkeit zum Anfrage-Tracklet.
4. Neusortierung der besten n Ergebnisse anhand einer Multi-Shot-Analyse (MSA) mit MaL dimensionsreduzierter Kovarianzdeskriptoren.

Alle Schritte müssen sowohl für jedes Anfrage-Tracklet als auch für die Galerie-Tracklets durchgeführt werden. Die Galerie-Tracklets können dabei allerdings vorberechnet werden, so dass während der Onlinephase die Schritte 1 und 2 nur für das Anfrage-Tracklet und neu in die Galerie hinzugefügte Tracklets durchgeführt werden müssen. Eine detaillierte Beschreibung des Verfahrens und der zugehörigen Algorithmen erfolgt in Abschnitt 7.2.

7.1 Herausforderungen bei der Wiedererkennung

An dieser Stelle werden die Herausforderungen aus Abschnitt 1.3 nochmal aufgegriffen und hinsichtlich der erscheinungsbasierten Personenwiedererkennung betrachtet. Die Aufgabe der Wiedererkennung von Personen im Videoüberwachungskontext ist eine sehr anspruchsvolle Aufgabe in der Bildauswertung, bei der sich insbesondere folgende Herausforderungen stellen:

- niedrige Auflösungen,
- stör- und kompressionsbedingte Bildartefakte (niedrige Bildqualität),
- weitere bildqualitätsbezogene Herausforderungen wie Zeilensprünge, Linsenverzerrungen etc.,
- Beleuchtungs- und Farbunterschiede,

Abbildung 7.3: Drei Beispiele für ähnliche Erscheinungen unterschiedlicher Personen: Die Bilder zeigen 6 verschiedene Personen.

- Verdeckungen,
- ähnliche Erscheinung unterschiedlicher Personen,
- unkontrollierte Umgebungen sowie
- unterschiedliche Kamerahersteller, -konfigurationen und -perspektiven.

Die meisten der aufgelisteten Schwierigkeiten stellen sich auch bei der Detektion und beim Tracking, wobei der Schwierigkeitsgrad dieser Herausforderungen bei der Wiedererkennung höher ist. Insbesondere die ähnlichen Erscheinungen stellen eine große Problematik dar (siehe Beispiele in Abbildung 7.3), die in der Regel mit der Anzahl der Galerie-Tracklets steigt. Deshalb sind diskriminative Personenrepräsentationen, gerade wenn sie handentworfen erstellt werden, eine wichtige Komponente, weshalb auch für diese Aufgabe Kovarianzdeskriptoren geeignet erscheinen. Kovarianzdeskriptoren wurden das erste Mal in [Bak10] für die erscheinungsbasierte Personenwiedererkennung eingesetzt [Bak14].

Unterschiedliche Kameratypen und -konfigurationen sind bei der Wiedererkennung auch problematischer als bei den beiden anderen betrachteten Bildauswerteaufgaben, da die Galerie-Tracklets sehr oft von vielen unterschiedlichen Kameras stammen. Im Gegensatz zur Detektion, bei der das Detektionsverfahren im Allgemeinen mit vielen Ansichten trainiert wird, und zum Tracking, bei dem das erscheinungsbasierte Modell über die Zeit

angepasst wird, muss bei der Wiedererkennung der Vergleich auch mit unterschiedlichen Perspektiven durchgeführt werden. Im ungünstigsten Fall erfolgt ein Vergleich der Vorderseite einer Person mit deren Rückseite, bei der die Zugehörigkeit der beiden Ansichten zu derselben Person nicht zuverlässig festgestellt werden kann.

Bei *kleineren* Unterschieden zwischen den Perspektiven bzw. Pose von Personen kann eine poseninvariante Repräsentation von Vorteil sein. Durch Nichtverwendung der x-Koordinate des Merkmalsvektors kann beispielsweise die Invarianz gegenüber der Pose einfach erhöht werden (vgl. [Hir11]), was im Bereich der Videoüberwachung einen großen Vorteil erzeugt. Durch diesen einfachen Schritt wird den Variationsunterschieden in der x-Richtung im Bild begegnet, die, im Gegensatz zu Variationsunterschieden in der y-Richtung, sehr häufig auftreten. Eine Skalierungsinvarianz kann bei der Wiedererkennung — im Vergleich zur Detektion — als gegeben erachtet werden, da als Eingabe in die Wiedererkennungsverfahren Bildausschnitte dienen, die durch die Person ausgefüllt sind.

Auch Farbunterschiede zwischen unterschiedlichen Kameras können bei der erscheinungsbasierten Wiedererkennung eine größere Herausforderung im Vergleich zu Detektions- und Trackingaufgaben darstellen. Während beispielsweise beim Tracking Beleuchtungs- und Farbunterschiede schritthaltend mit der Bildwiederholrate kompensiert werden können, in dem z.B. die Personenrepräsentation nach jedem neuen Bild angepasst wird, können bei der Wiedererkennung zwei Tracklets mit drastischen Beleuchtungs- und Farbunterschieden vorliegen, die miteinander verglichen werden müssen. Deshalb erscheint aufgrund der guten Eigenschaften hinsichtlich Invarianz gegenüber mittelwertbasierten Verschiebungen von Helligkeitswerten etc. der Kovarianzdeskriptor als geeigneter Deskriptor für diese Aufgabe (vgl. Kapitel 3). Andernfalls müssten die Kameras gegenüber beispielsweise Farbunterschieden durch z.B. Farbkalibrierung der Kameras aufeinander abgestimmt werden.

Eine weitere Problematik bei der trackletbasierten Wiedererkennung ist die Abhängigkeit von einem Trackingverfahren, das in der Regel zur Tracklet-Erzeugung vor die Wiedererkennung geschaltet wird. Die Erzeugung erfolgt im Allgemeinen unkontrolliert, d.h. ohne Verifizierung der ausgestanzten Personen. Dies kann dazu führen, dass in einigen Fällen eine Person nicht

zentriert oder nur teilweise in einem Bildausschnitt sichtbar ist. Auch kann es passieren, dass mehr als eine Person in einem Bildausschnitt sichtbar ist. In allen drei genannten Fällen wird ein erscheinungsbasiertes Wiedererkennungsverfahren, ohne implizite Detektion dieser Fälle, davon ausgehen, dass nur eine Person im Bildausschnitt zu sehen ist. Detektionsverfahren, die beispielsweise diese Fälle explizit erkennen können, würden die Wiedererkennung robuster machen. Das vorgestellten Wiedererkennungsverfahren kann solche *Ausreißer* während der Tracklet-Deskriptor-Berechnung eliminieren.

Eine weitere Schwierigkeit bei erscheinungsbasierten Wiedererkennungsverfahren auf Basis von bildregionenbasierten Deskriptoren ist der *Hintergund*, falls dieser in die Personenrepräsentation mit einfließt. Dazu können beispielsweise Segmentierungsverfahren, wie z.B. der Ansatz in [Mon13], zur Entfernung des Hintergrunds vor Wiedererkennungsverfahren geschaltet werden. Je weniger Hintergrund in Bildausschnitten sichtbar ist, desto weniger *Störeinflüsse* liegen vor, woraus im Allgemeinen eine bessere Wiedererkennungsleistung resultiert. Im Bereich der Videoüberwachung werden die Hintergründe aufgrund der in der Regel vorliegenden niedrigen Videoqualität oft nur unzureichend entfernt (siehe Abbildung 3.4). Das vorgestellte Wiedererkennungsverfahren begegnet den Segmentierungsfehlern durch den mittelwertbasierten Ansatz der Tracklet-Deskriptor-Berechnung.

7.2 Das erscheinungsbasierte Wiedererkennungsverfahren KovIDent

In diesem Abschnitt wird das erscheinungsbasierte Verfahren zur Wiedererkennung von Personen, das auf dem mathematischen Rahmenwerk in Kapitel 4 beruht, im Detail vorgestellt. Das Verfahren wird im Folgenden als *kovarianzdeskriptorbasiertes Identifikationsverfahren* (KovIDent) bezeichnet.

Seien $\boldsymbol{T}_a$ ein Anfrage-Tracklet einer gesuchten Person und $\mathcal{G} = \{\boldsymbol{T}_g\}$, $g = 1, \ldots, n$ die Menge aller Galerie-Tracklets, mit denen $\boldsymbol{T}_a$ verglichen werden soll. Für alle $\boldsymbol{T}_g$ wurden vorab Tracklet-Deskriptoren berechnet und mit diesen gespeichert. Für den aktuellen Vergleich wird

nun auch für $\boldsymbol{T}_a$ der Tracklet-Deskriptor berechnet, wozu auch für alle Rechtecke von $\boldsymbol{T}_a$ die hierfür notwendigen Merkmale zu berechnen sind.

7.2.1 Kovarianzdeskriptor für Einzelbilder

Bis auf die Einschränkung, dass die Merkmale als Skalare oder Vektoren mit reellen Zahlen vorliegen müssen, können für das Wiedererkennungsverfahren KovIDent beliebige Merkmale verwendet werden. Allerdings sollten bei der Auswahl der Merkmale anwendungsspezifische Faktoren, wie beispielsweise die Auflösung der Bilder etc., mitberücksichtigt werden. Bei niedrig aufgelösten Bildern können biometrische oder semantische Merkmale, wie beispielsweise explizite Detektionen von Accessoires an der Person, nur bedingt extrahiert werden, da in der Regel die Auflösung dafür zu niedrig ist (vgl. Kapitel 3).

Bei niedrig aufgelösten Personen bieten sich für erscheinungsbasierte Wiedererkennungsverfahren einfache Merkmale an. Insbesondere Farbmerkmale spielen dabei eine wichtige Rolle, da die Farbe der Kleidung etc. oft der einzige Hinweis auf die Identität der Person ist. Je nach Auflösung können zusätzlich auch z.B. gradientenbasierte Merkmale verwendet werden, mit denen eine Wiedererkennung von Texturen oder Mustern auf der Kleidung möglich wird [Sag14]. Voraussetzung dafür sind allerdings ausreichend hoch aufgelöste Bilder, damit genügend solcher Merkmale berechnet werden können. Gradientenbasierte Merkmale können auch für einen Vergleich von Körperkonturen verwendet werden. Aufgrund der hohen Intra-Varianz von Personenkonturen können solche Merkmale die Wiedererkennungsleistung allerdings auch verschlechtern, weswegen sie vielmehr für biometrische Ansätze verwendet werden, die versuchen, Personen anhand ihrer Gangart wiederzuerkennen [Kaw12, Cho13].

Für den im Rahmen dieser Arbeit erarbeiteten Ansatz wurden niedrigdimensionale Merkmalsvektoren definiert, um eine möglichst anwendungsunabhängige Wiedererkennung durchführen zu können. Das Verfahren verwendet einfache Merkmale: Farbwerte der Pixel (RGB-Farbraum) und die y-Bildkoordinate. Die x-Koordinate wird nicht betrachtet, da sie die Intra-Varianz für Fälle, in denen eine Person in unterschiedlichen Seitenansichten zu sehen ist, erhöht und somit im Allgemeinen die Wiedererken-

nungsleistung senkt (vgl. [Hir11]). Der Einfluss von mehreren Farbräumen hinsichtlich Wiedererkennungsleistung wird in Kapitel 8 untersucht.

Der Kovarianzdeskriptor für eine ausgestanzte Person bestimmt sich somit aus den y-Pixelkoordinaten und den RGB-Pixelwerten. Ein Merkmalsvektor $\boldsymbol{f}_{(x,y)}$ für die Koordinate (x, y) bzgl. der linken oberen Ecke eines Bildausschnitts $\boldsymbol{I_R}$ mit der Koordinate $(0, 0)$ ist wie folgt definiert:

$$\boldsymbol{f}_{(x,y)} = \begin{pmatrix} Y(x,y) \\ R(x,y) \\ G(x,y) \\ B(x,y) \end{pmatrix} , \tag{7.1}$$

mit $Y(x, y) := y$.

Die Kovarianzdeskriptoren berechnen sich wieder gemäß Abschnitt 3.4.1.

7.2.2 Tracklet-Deskriptor

Im zweiten Verfahrensschritt werden Tracklet-Deskriptoren für die Repräsentation der Tracklets berechnet. Dazu werden aus n_t Kovarianzdeskriptoren eines Tracklets $\tilde{t}_c$ Tracklet-Deskriptoren bestimmt, um mehrere Ansichten etc. zu berücksichtigen. Die Berechnung sowohl der Tracklet-Deskriptoren als auch deren Anzahl $\tilde{t}_c$ für ein Tracklet erfolgt auf der Basis des Spectral Clusterings [vL07], was dem Prinzip des lokalen MaL Verfahrens Laplacian Eigenmaps ähnelt (vgl. Abschnitt 4.2.1). Der Ansatz wird im Folgenden ausführlich vorgestellt.

Seien $\{\boldsymbol{\Sigma}_{\boldsymbol{R}_i}\}$, $i = 1, \ldots, n_t$ die Kovarianzdeskriptoren, für die Tracklet-Deskriptoren berechnet werden sollen. Zunächst wird ein ungerichteter, gewichteter Adjazenzgraph $\boldsymbol{W} = (w_{i,j})$ mit der folgenden zugehörigen symmetrischen Matrix aufgebaut:

$$w_{i,j} = w_{j,i} = \begin{cases} 1\,, & \text{wenn } \boldsymbol{\Sigma}_{\boldsymbol{R}_j} \in N_i \\ 0\,, & \text{sonst}\,. \end{cases} \tag{7.2}$$

$\boldsymbol{\Sigma}_{\boldsymbol{R}_j}$ ist Nachbar von $\boldsymbol{\Sigma}_{\boldsymbol{R}_i}$ ($\boldsymbol{\Sigma}_{\boldsymbol{R}_j} \in N_i$), wenn der geodätische Abstand zwischen den beiden Kovarianzdeskriptoren kleiner einem Schwellwert ϵ ist. Der geodätische Abstand berechnet sich durch die Gleichung (4.8).

Alternative Ansätze bestimmen die Nachbarschaftsbeziehungen durch die k nächsten Nachbarn für jeden Deskriptor oder es werden alle Deskriptoren miteinander verbunden. Zusätzlich können bei allen Ansätzen die Nachbarschaftsverbindungen gewichtet werden. Inwieweit die Auswahl der Methode zur Bestimmung der Nachbarschaftsbeziehung das Spectral Clustering beeinflusst, ist im Allgemeinen nicht bekannt [vL07].

Anschließend wird ausgehend von dem Adjazenzgraphen folgende normalisierte $n_t \times n_t$ Laplacian Matrix berechnet:

$$\boldsymbol{L'} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{D}^{-\frac{1}{2}}\,, \tag{7.3}$$

wobei $\boldsymbol{D} = (d_{i,i})$ die durch die Gleichung (4.21) definierte Knotengradmatrix ist.

$\boldsymbol{L'}$ hat folgende, für die Arbeit relevante Eigenschaften (vgl. [vL07]):

- 0 ist ein Eigenwert von $\boldsymbol{L'}$ (mit dem Eigenvektor $\boldsymbol{D}^{\frac{1}{2}}\,\vec{\mathbf{1}}$).
- $\boldsymbol{L'}$ ist positiv semidefinit und hat n nichtnegative reelle Eigenwerte $0 = \lambda_1 \leq \cdots \leq \lambda_n$.
- Die Multiplizität des Eigenwerts 0 von $\boldsymbol{L'}$ entspricht der Anzahl von maximal zusammenhängenden Teilgraphen (Zusammenhangskomponenten) des (ggf. unzusammenhängenden) Adjazenzgraphen.
- O.B.d.A hat $\boldsymbol{L'}$ eine Blockdiagonalform, unter der Annahme, dass die zu einer Zusammenhangskomponente $\boldsymbol{K}_i$, $i = 1, \ldots, t_c$ gehörenden Kovarianzdeskriptoren durch aufeinanderfolgende Zeilen von $\boldsymbol{L'}$ repräsentiert werden:

$$\boldsymbol{L'} = \begin{pmatrix} \boldsymbol{K}_1 & & & \\ & \boldsymbol{K}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{K}_{t_c} \end{pmatrix}, \tag{7.4}$$

  wobei t_c der Anzahl der Zusammenhangskomponenten entspricht.
- Das Eigenwertspektrum von $\boldsymbol{L'}$ entspricht der Vereinigung der Eigenwertspektren von $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_{t_c}$.

- Die Eigenvektoren von $\boldsymbol{L'}$ sind jeweils einer Zusammenhangskomponente zugeordnet und haben an Stellen anderer Zusammenhangskomponenten Null-Einträge.

Bestimmung der maximalen Anzahl von Tracklet-Deskriptoren

Vor der Berechnung der Tracklet-Deskriptoren muss deren maximale Anzahl bestimmt werden. Die exakte Anzahl t_c wird im darauffolgenden Berechnungsschritt ermittelt, während der Berechnung der Tracklet-Deskriptoren.

Die maximale Anzahl der Tracklet-Deskriptoren wird so bestimmt, dass sie der Anzahl der Zusammenhangskomponenten des Graphen entspricht, der durch $\boldsymbol{L'}$ repräsentiert wird. Im Idealfall hat der Graph t_c Zusammenhangskomponenten $\boldsymbol{K}_i$, $i = 1 \dots t_c$, die nicht zusammenhängen:

$$\boldsymbol{L'} = \begin{pmatrix} \boldsymbol{K}_1 & & & \\ & \boldsymbol{K}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{K}_{t_c} \end{pmatrix}, \tag{7.5}$$

was der Multiplizität des Eigenwerts 0 von $\boldsymbol{L'}$ entspricht. In der Praxis sind in der Regel die Zusammenhangskomponenten allerdings nur selten unzusammenhängend, d.h. es gibt gemäß der Graphentheorie oft nur eine Zusammenhangskomponente. Ein einzelner Tracklet-Deskriptor ist im Allgemeinen aber nicht ausreichend für eine Repräsentation eines Tracklets, da dann ggf. verschiedene Ansichten ungünstig durch einen einzelnen Deskriptor repräsentiert und Ausreißer unzureichend behandelt werden.

Im Rahmen dieser Arbeit werden zur Bestimmung der maximalen Anzahl von Tracklet-Deskriptoren *schwach zusammenhängende* Zusammenhangskomponenten ermittelt. Dies ist allerdings nicht so einfach wie unzusammenhängende Komponenten zu identifizieren, da die Anzahl der schwach zusammenhängenden Komponenten nicht anhand der Multiplizität des Eigenwerts 0 von $\boldsymbol{L'}$ bestimmt werden kann. Die Bestimmung der schwach zusammenhängenden Komponenten erfolgt mittels einer Analyse von Eigenwert-Differenzen benachbarter Eigenwerte von $\boldsymbol{L'}$, d.h. anhand

der Differenzen zwischen λ_{t_c} und λ_{t_c+1}, für $t_c = 1, \ldots, n-1$, wobei die Eigenwerte aufsteigend sortiert sind: $\lambda_1 \leq \cdots \leq \lambda_n$. Die Eigenwert-Differenzen werden mit einem Schwellwert δ verglichen, beginnend mit der Differenz für $t_c = 1$. Sobald $\lambda_{t_c+1} - \lambda_{t_c} > \delta$ wird der Algorithmus abgebrochen. Der Wert des Indexes t_c bei Abbruch des Algorithmus gibt die maximale Anzahl von Tracklet-Deskriptoren an.

Berechnung der Tracklet-Deskriptoren

Für jede Zusammenhangskomponente $\boldsymbol{K}_i$, $i = 1, \ldots, t_c$ wird ein Tracklet-Deskriptor berechnet, welcher dem Mittelwert der Kovarianzdeskriptoren entspricht, die der Zusammenhangskomponente $\boldsymbol{K}_i$ zugehören. Die Zugehörigkeit wird mit einem partitionierenden Verfahren ermittelt. Dazu werden zunächst die Eigenvektoren von $\boldsymbol{L'}$ zu den t_c kleinsten Eigenwerten berechnet und spaltenweise durch eine Matrix repräsentiert. Anschließend werden die Zeilen der Matrix mit dem *k-Means*-Algorithmus [Har79] geclustert. Nach dem Clustern wird zudem eine Überprüfung der Clustergrößen durchgeführt. Ist die Anzahl der Kovarianzdeskriptoren eines Clusters kleiner als ein vorgegebener Schwellwert, wird dieses Cluster entfernt, so dass für die endgültige Anzahl der Tracklet-Deskriptoren $\tilde{t}_c \leq t_c$ gilt. Bei einem zu groß gewählten Schwellwert oder einem Tracklet, bei dem sich die Erscheinung der Person stark von Bild zu Bild ändert, ist auch die Anzahl $\tilde{t}_c = 0$ möglich. Eine Clustergröße von fünf hat sich als geeignet erwiesen. Fälle, in denen ausschließlich kleine Cluster erzeugt werden, deren Größe kleiner als der Schwellwert ist, traten in den durchgeführten Evaluationen nicht auf. Deshalb wird in dieser Arbeit nicht weiter darauf eingegangen.

Anschließend wird für jedes der $\tilde{t}_c$ übriggebliebenen Cluster $\boldsymbol{C}_i$, $i = 1, \ldots, \tilde{t}_c$ ein empirischer Mittelwert aus den Kovarianzdeskriptoren des jeweiligen Clusters berechnet. Seien $\{\boldsymbol{\Sigma}_{\boldsymbol{C}_i j}\}$, $j = 1, \ldots, n$ die Kovarianzdeskriptoren des Clusters $\boldsymbol{C}_i$, dann berechnet sich der Mittelwert $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}$ für das Cluster $\boldsymbol{C}_i$ gemäß Abschnitt 4.1.2: Der Mittelwert wird iterativ mit dem Gauß-Newton-Verfahren bestimmt, wobei ein Iterationsschritt durch die Gleichung (4.10) definiert ist:

$$\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}^{t+1} = \exp_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}^{t}} \left(\frac{1}{n} \sum_{j=1}^{n} \log_{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}^{t}} (\boldsymbol{\Sigma}_{\boldsymbol{C}_i j}) \right) . \tag{7.6}$$

Der Iterationsschritt wird solange wiederholt, bis die Bedingung $g(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}^{t+1}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}^{t}) < \psi$ erfüllt ist. Im ersten Iterationsschritt ($t = 0$) wird ein beliebiger Kovarianzdeskriptor aus der Menge $\{\boldsymbol{\Sigma}_{\boldsymbol{C}_i j}\}$, $j = 1, \ldots, n$ für $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}^{0}$ eingesetzt.

Ein Tracklet $\boldsymbol{T}$ wird somit durch die Menge der Mittelwerte $\{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}\}$, $i = 1, \ldots, \tilde{t}_c$ repräsentiert.

7.2.3 Tracklet-Vergleich

Der Vergleich von Tracklets erfolgt mittels Vergleich von Tracklet-Deskriptoren. Ein großer Vorteil bei diesem Ansatz ist, dass ein Tracklet durch eine geringe Anzahl an Deskriptoren repräsentiert wird und somit der Vergleich effizienter durchgeführt werden kann.

Die Durchführung einer MSA, bei der alle Einzelbilder aller Tracklets miteinander verglichen werden, wäre eine Alternative zu dem hier vorgestellten trackletbasierten Ansatz. Allerdings ist eine MSA bei großen Videodatenbanken im Allgemeinen unzweckmäßig, da der Vergleich eines Anfrage-Tracklets mit den Galerie-Tracklets zu lange dauert [Her18]. Eine einfache Skalierbarkeit könnte beispielsweise durch den Vergleich von zufällig ausgewählten Einzelbildern anstatt aller Einzelbilder erreicht werden. Dabei kann allerdings nicht garantiert werden, dass die Aufnahmen einer Person, die für den Vergleich sehr gut geeignet sind, miteinander verglichen werden. Im ungünstigsten Fall werden ausschließlich Einzelbilder verglichen, die z.B. Störungen durch Bildartefakte aufweisen oder die Person unvollständig abgebildet oder aus einem ungeeigneten Winkel zu sehen ist.

Im Vergleich dazu bietet der trackletbasierte Ansatz einen weiteren Vorteil. Durch die Mittelwertberechnung wird deutlich die Gefahr gemindert, die bei der MSA besteht, dass zwei *Ausreißer*, die unterschiedliche Personen zeigen, aufgrund z.B. geringer Bildqualität den geringsten Abstand zueinander (höchste Ähnlichkeit) aufweisen.

Seien die Mengen $\{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}\}, i = 1, \ldots, \tilde{t}_{c_i}$ und $\{\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C'}_j}\}, j = 1, \ldots, \tilde{t}_{c_j}$ Tracklet-Deskriptoren für Tracklet $\boldsymbol{T}$ bzw. Tracklet $\boldsymbol{T'}$. Die Berechnung der Ähnlichkeit der Tracklets erfolgt anhand eines paarweisen Vergleichs der einzelnen

$\boldsymbol{T}$ zugehörigen Tracklet-Deskriptoren mit denen aus $\boldsymbol{T}'$. Dazu werden mittels der Gleichung (4.8) die geodätischen Abstände zwischen den Tracklet-Deskriptoren berechnet. Die Ähnlichkeit zwischen $\boldsymbol{T}$ und $\boldsymbol{T}'$ entspricht dem kleinsten geodätischen Abstand

$$\min_{i,j} \mathrm{g}(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}_i}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{C}'_j}), \quad i = 1, \ldots, \tilde{t}_{c_i}\,, \quad j = 1, \ldots, \tilde{t}_{c_j}\,. \tag{7.7}$$

Bei einem Vergleich eines Anfrage-Tracklets mit mehreren Galerie-Tracklets werden die Ergebnisse anhand ihrer Ähnlichkeiten zum Anfrage-Tracklet sortiert und anschließend aufgelistet.

7.2.4 Einzelbildbasierte Neusortierung

Bei der einzelbildbasierten Neusortierung werden die besten z Ergebnisse (Galerie-Tracklets) aus der Liste des Tracklet-Vergleichs anhand einer MSA neu geordnet, wobei die Analyse in der Regel aus Zeitgründen nur mit den ersten z Ergebnissen erfolgt. Die Wahl von z ist von dem Einsatzszenario (erforderliche maximale Anfragezeit), der verfügbaren Rechenleistung etc. abhängig. Bei ausreichend Ressourcen kann die MSA theoretisch auf die gesamte Galerie angewandt werden.

Die MSA erfolgt anhand eines Vergleichs der einzelnen Kovarianzdeskriptoren. Es wird — ähnlich wie in Abschnitt 5.2 — die Annahme getroffen, dass die Deskriptoren einer Person auf oder nahe bei einer (möglicherweise nichtlinearen) Mannigfaltigkeit liegen, die im Sym_n^+ eingebettet ist. Außerdem wird — ähnlich wie bei der Körperteildetektion — erwartet, dass sich Untermannigfaltigkeiten verschiedener Personen unterscheiden lassen.

Im ersten Schritt der MSA wird die Gesamtmannigfaltigkeit aller Kovarianzdeskriptoren der ersten z Galerie-Tracklets berechnet. Dies erfolgt mittels dem Laplacian Eigenmaps Verfahren, das in Abschnitt 4.2.1 beschrieben ist. Das Verfahren verfolgt einen lokalen Ansatz, der im Gegensatz zu einer globalen Strategie im Allgemeinen besser für ein breiteres Spektrum an Mannigfaltigkeiten geeignet ist.

Sei $\mathcal{A} = \{\boldsymbol{\Sigma}_i\}$, $i = 1, \ldots, n$ die Menge der Kovarianzdeskriptoren, die den ersten z Tracklet-Ergebnissen zugrunde liegen, und der Kovarianzdeskriptoren des Anfrage-Tracklets. Die niedrigdimensionale (d-dimensionale)

Gesamtrepräsentation $\boldsymbol{Y}$ der Untermannigfaltigkeiten einzelner Tracklets ergibt sich aus den Eigenvektoren $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ der normalisierten Laplace-Matrix $\boldsymbol{L}'$, die zu den kleinsten Eigenwerten $\lambda_i \neq 0 \leq \cdots \leq \lambda_{i+d}$ von $\boldsymbol{L}'$ gehören:

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)^T \ . \tag{7.8}$$

Die Dimension d wird dabei in Abhängigkeit der Anzahl der Merkmale fest vorgegeben ($d <$ Anzahl Merkmale) und muss für die MSA nicht zwingend der wahren intrinsischen Dimension der Gesamtmannigfaltigkeit entsprechen [Vur15]. $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ entspricht dann der d-dimensionalen Einbettung der Kovarianzdeskriptoren aus $\mathcal{A}$, wobei die erste Spalte von $\boldsymbol{Y} = \boldsymbol{Y}(:, 1)$ den Kovarianzdeskriptor $\boldsymbol{\Sigma}_1$ und die letzte Spalte den Kovarianzdeskriptor $\boldsymbol{\Sigma}_n$ repräsentiert.

Sei $\boldsymbol{Y}_a$ die Repräsentation des Anfrage-Tracklets $\boldsymbol{T}_a$ und $\boldsymbol{Y}_z$ die Repräsentation der ersten z Ergebnisse. Im ersten Schritt der MSA wird dann zunächst für jeden Punkt aus $\boldsymbol{Y}_a$ mittels des euklidischen Abstands der nächste Nachbar aus $\boldsymbol{Y}_z$ gesucht. Anschließend wird für jedes Tracklet der Anteil an *nächsten Punkten* zum Anfrage-Tracklet bestimmt. Die Neusortierung der z Galerie-Tracklets $\{\mathcal{T}_i\}$, $i = 1, \ldots, z,$ erfolgt anhand des Ähnlichkeitswerts

$$d_a = \frac{\text{Anzahl der nächsten Nachbarn von } \boldsymbol{T}_i}{\text{Anzahl der Deskriptoren von } \boldsymbol{T}_a}\,, \tag{7.9}$$

wobei 1 eine hohe und 0 eine niedrige Ähnlichkeit bedeutet.

7.3 Verfahrensevaluation

KovIDent wurde anhand von zwei eigenen Tracklet-Datensätzen evaluiert. Die erste Evaluation erfolgte auf einem Kameranetzwerk-Datensatz, der ein typisches Videoüberwachungsszenario abbildet. Die Aufzeichnungen zeigen sehr niedrig aufgelöste Personen, die aus einer Schrägsicht in einer öffentlichen Halle akquiriert wurden. Für die zweite Evaluation wurde ein Fahndungsdatensatz betrachtet, der im Kontext der forensischen Personensuche erstellt wurde.

Kameranetzwerk-Datensatz. Der Kameranetzwerk-Datensatz wurde im Rahmen des *CamInSens*[1]-Projekts in einer Eingangshalle[2] der Leibniz Universität Hannover akquiriert und besteht ausschließlich aus Innenaufnahmen. Er enthält 96 Tracklets von 10 unterschiedlichen Personen, zu denen jeweils mindestens zwei Tracklets existieren. Die Bildsequenzen stammen von drei verschiedenen Kameras (16 Sequenzen von Kamera A, 25 Sequenzen von Kamera B und 55 Sequenzen von Kamera C), deren Auflösung identisch sind (704 Pixel (Breite) $\times$ 576 Pixel (Höhe)). Die Kameras waren auf der gleichen Höhe (ca. 5 Meter) installiert und jeweils ca. 90° zueinander gedreht. Die unskalierten Auflösungen der einzelnen Personen sind maximal ca. 130 Pixel in der Höhe und unterscheiden sich aufgrund der Entfernung der Personen zur Kamera stark. Eine Bildsequenz besteht aus mindestens 10 und bis zu 1000 ausgestanzten rechteckigen Bildausschnitten einer Person, wobei die Bildausschnitte aller Sequenzen auf eine einheitliche Größe skaliert wurden: 64 Pixel (Breite) $\times$ 128 Pixel (Höhe). Außerdem sind die Personen während des Trackings mittels einem Segmentierungsverfahren anhand ihrer Bewegung grob vom Hintergrund getrennt worden, wobei der Hintergrund schwarz eingefärbt wurde. Abbildung 7.4 zeigt beispielhaft sechs Bildsequenzen aus dem Kameranetzwerk-Datensatz, von denen jeweils zwei Sequenzen von einer der drei Kameras stammen.

Fahndungsdatensatz. Der Fahndungsdatensatz wurde im Rahmen des MisPel[3]-Projekts auf dem Gelände der ehemaligen Karlsruher Parfümerie- und Toilettenseifenfabrik Wolff & Sohn[4] akquiriert und umfasst ausschließlich Außenaufnahmen. Er beinhaltet 41 Sequenzen von insgesamt 16 unterschiedlichen Personen, die aus derselben festen Kameraposition und -orientierung aufgezeichnet wurden. Die Bildauflösung beträgt 1920 Pixel in der Breite und 1080 Pixel in der Höhe. Für 6 Personen existiert

[1] `http://www.caminsens.org` CamInSens - *Verteilte vernetzte Kamerasysteme zur in situ-Erkennung Personen-induzierter Gefahrensituationen.* BMBF-Forschungsprojekt mit Förderkennzeichen 13N10809 - 13N10814.

[2] `https://www.uni-hannover.de/fileadmin/luh/images/webredaktion/service/wegweiser/lageplan/lageplan.jpg` (Gebäude 1101)

[3] `https://www.sifo.de/de/mispel-multi-biometriebasierte-forensische-personensuche-in-lichtbild-und-videomassendaten-2105.html` MisPel - *Multi-Biometriebasierte Forensische Personensuche in Lichtbild- und Videomassendaten.* BMBF-Forschungsprojekt mit Förderkennzeichen 13N12059 - 13N12065.

[4] `http://stadtlexikon.karlsruhe.de/index.php/De:Lexikon:ins-0041`

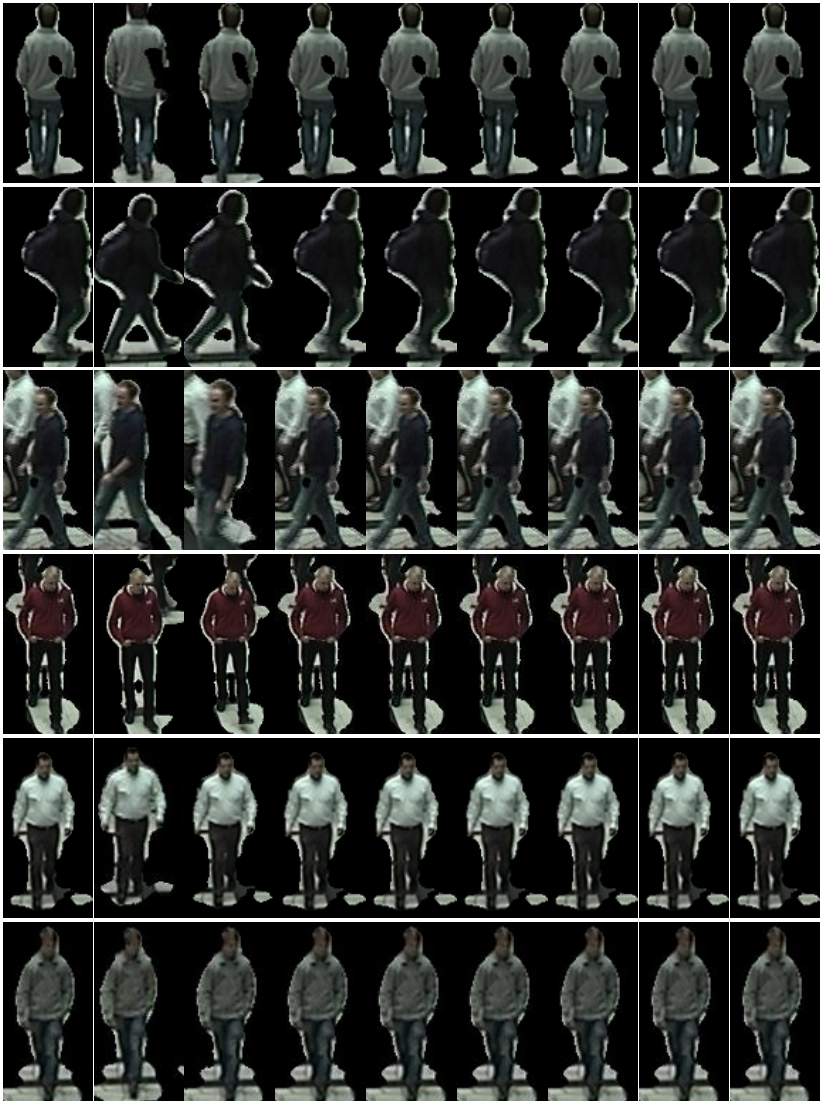


Abbildung 7.4: Beispiel-Tracklet-Ausschnitte sechs unterschiedlicher Personen aus dem Kameranetzwerk-Datensatz.

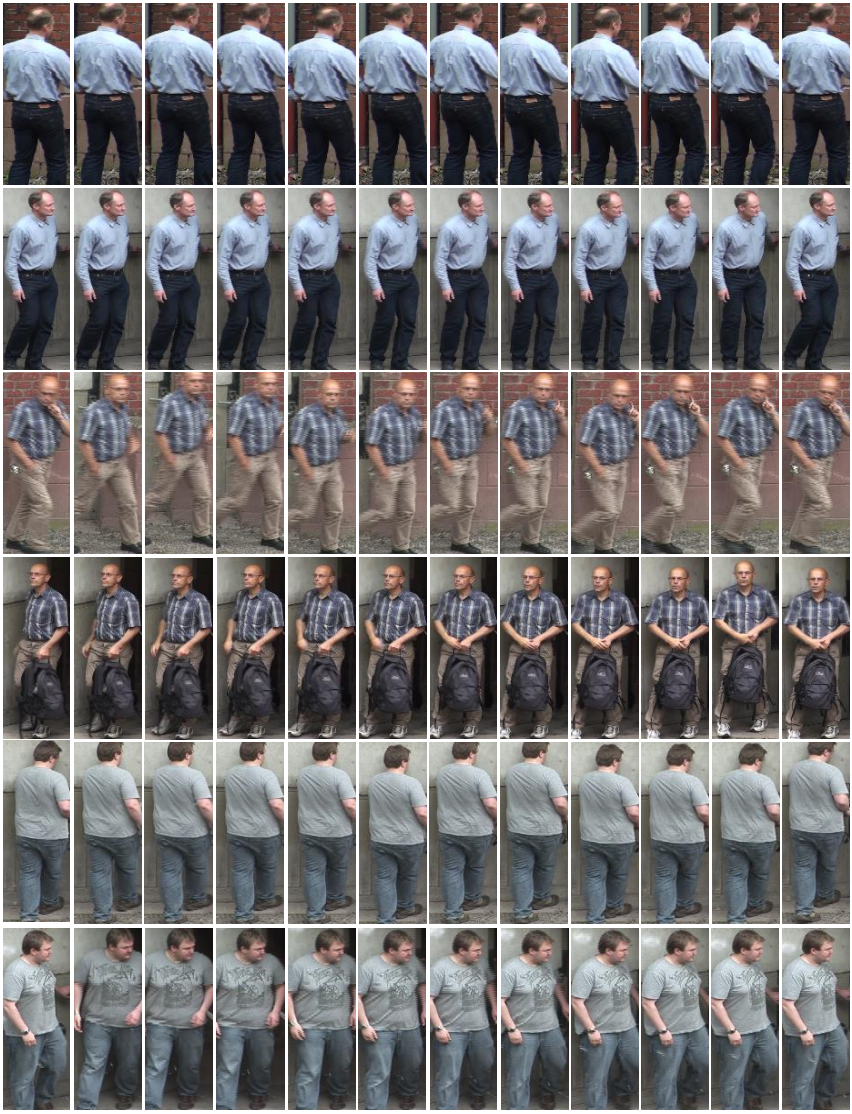


Abbildung 7.5: Beispiel-Tracklet-Ausschnitte drei unterschiedlicher Personen aus dem Fahndungsdatensatz.

jeweils nur eine Sequenz, weswegen bei der Evaluation nur die 10 Personen berücksichtigt wurden, zu denen es mehrere Sequenzen gibt. Die Auflösungen der Personen in diesem Datensatz sind deutlich höher als beim Kameranetzwerk-Datensatz und variieren zudem stärker. Die Höhen einer Person in Pixel liegen zwischen ca. 600 und 850 Pixel, so dass auch biometrische Verfahrensansätze zur Personenwiedererkennung eingesetzt werden können. Für die Evaluation von KovIDent wurden die Bildausschnitte auf eine einheitliche Höhe von 700 Pixel skaliert, um ein Evaluation nah an der Originalauflösung durchführen zu können. Aufgrund der starken Variationen in den Auflösungen wurden die Seitenverhältnisse bei der Skalierung beibehalten. Beim Fahndungsdatensatz besteht eine Bildsequenz aus mindestens 13 und bis zu 2285 ausgestanzten rechteckigen Bildausschnitten einer Person. Im Gegensatz zum Kameranetzwerk-Datensatz sind die Personen beim Fahndungsdatensatz nicht segmentiert. Abbildung 7.5 zeigt einige Ausschnitte des Fahndungsdatensatzes.

7.3.1 Evaluation mit dem Kameranetzwerk-Datensatz

Bei dieser Evaluation wurde je ein Tracklet aus der Galerie genommen (Anfrage-Tracklet) und mit den restlichen Galerie-Tracklets verglichen. Die Evaluation wurde unter *erschwerten* Bedingungen durchgeführt: Wenn es neben dem Anfrage-Tracklet mehrere Galerie-Tracklets zu einer Person gab, wurden bei einer Suchanfrage mit dieser Person nur ein Galerie-Tracklet in der Galerie gelassen. Existieren neben dem Anfrage-Tracklet z.B. drei zugehörige Galerie-Tracklets, wurden drei Durchläufe mit jeweils einem dieser Tracklets durchgeführt.

Die Evaluation wurde in zwei Teilen durchgeführt. Zunächst wurde das Ergebnis von KovIDent nach Schritt 3 (Vergleich des Anfrage-Tracklets mit den Galerie-Tracklets) betrachtet, d.h. ohne Neusortierung der Ergebnisliste. Dieses Ergebnis wurde außerdem mit einem Ansatz verglichen, der Farbhistogramme zur Personenrepräsentation verwendet. Bei dem Histogramm-Ansatz wird eine MSA durchgeführt. Für jedes ausgestanzte Personenbild eines Anfrage-Tracklets werden RGB-Histogramme mit 64 Bins berechnet und mit den Histogrammen der Galeriesequenzen verglichen. Bei dem Vergleich wird der kleinste Bhattacharya-Abstand aller

paarweisen Abstände der Anfrage- und Galerie-Histogrammen ermittelt, anhand dessen die Ergebnisliste sortiert wird.

Im zweiten Teil wurde KovIDent bis Schritt 4 (Neusortierung der besten z Ergebnisse) evaluiert. Neben der in Abschnitt 7.2.4 beschriebenen Neusortierung, die auf der Einbettung der Kovarianzdeskriptoren basiert, wurde zum Vergleich eine Neusortierung durchgeführt, bei der analog zur Evaluation des Histogramm-Ansatzes eine MSA durchgeführt wurde, nur mit dem Unterschied, dass die paarweisen geodätischen Abstände zwischen den Kovarianzdeskriptoren betrachtet wurden (Referenz-Neusortierung). Im Rahmen des Videoüberwachungsszenerios, in dem die Personen automatisch kameraübergreifend zugeordnet werden sollen, sind ausschließlich die oberen Ränge von Bedeutung, weshalb die ersten 5 Ränge in Betracht gezogen worden.

Das Evaluationsergebnis ist durch die mittlere RPR angegeben, die das durchschnittliche Verhältnis der gefundenen Personen unter den ersten x Einträgen in der — nach der Ähnlichkeit sortierten — Trefferliste zur Gesamtzahl der *relevanten* Personen in der Galerie angibt. Da bei den Suchanfragen im Rahmen der Evaluationen jeweils nur eine *relevante* Person in der Galerie vorhanden war, ist das Verhältnis einer Suchanfrage 0 oder 1. Die mittleren RPRn des Histogramm-Ansatzes, von KovIDent nach Schritt 3 sowie die Ergebnisse von KovIDent mit den beiden jeweiligen Neusortierungen (Schritt 4) sind in der Tabelle 7.1 zusammengefasst.

Obwohl die Histogramme und Kovarianzdeskriptoren auf ähnliche Merkmalskombinationen basieren, erzielte das KovIDent-Verfahren gegenüber dem Referenzverfahren (Histogramm-Ansatz) deutlich bessere RPRn. Das resultiert, neben den in Kapitel 3 aufgeführten Vorteilen der Kovarianzdeskriptoren, aus der Strategie, die Tracklets durch Mittelwerte zu repräsentieren.

Die beiden Ansätze zur Neusortierung konnten die KovIDent-Ergebnisse nach Schritt 4 weiter verbessern. Lediglich die Referenz-Neusortierung auf Basis der geodätischen Abstände führte — wahrscheinlich aufgrund von Ausreißern — zu einer geringeren Genauigkeit für den ersten Rang. Die Neusortierung in der *Einbettung* lieferte für alle Ränge bessere Ergebnisse als die Referenz-Neusortierung. In einer weiteren Evaluation wurden die Neusortierungen auf einem Fahndungsdatensatz evaluiert. Die Ergebnisse werden im folgenden Abschnitt aufgeführt.

Rang	**Histogramm-Ansatz**	**KovIDent** Schritt 3	**KovIDent** Schritt 4 geodätisch	**KovIDent** Schritt 4 eingebettet
1	0,389	**0,646**	0,588	**0,673**
≤ 2	0,473	**0,715**	0,72	**0,737**
≤ 3	0,551	**0,767**	0,794	**0,821**
≤ 4	0,626	**0,801**	0,827	**0,839**
≤ 5	0,643	**0,839**	**0,839**	**0,839**

Tabelle 7.1: Mittlere RPRn der Wiedererkennungsverfahren aus der Evaluation mit dem Kameranetzwerk-Datensatz. Die Rang x bezogenen mittleren RPRn geben den Anteil der korrekt gefundenen Personen unter den ersten x Einträgen der Trefferliste an. In der zweiten Spalte von rechts ist das KovIDent-Ergebnis nach der Neusortierung mittels der paarweisen geodätischen Abständen (Referenz-Neusortierung) aufgeführt und in der rechten Spalte das KovIDent-Ergebnis nach der Neusortierung in der Einbettung. Die Neusortierungen wurden jeweils unter den ersten 5 Rängen durchgeführt. Die besten Ergebnisse sind fett hervorgehoben.

Rang	**KovIDent** Schritt 3	**KovIDent** Schritt 4 geodätisch	**KovIDent** Schritt 4 eingebettet
1	0,445	0,195	**0,719**
≤ 2	0,539	0,281	**0,719**
≤ 3	0,563	0,367	**0,727**
≤ 4	0,633	0,398	**0,727**
≤ 5	0,664	0,422	**0,727**
≤ 10	0,727	0,586	**0,734**

Tabelle 7.2: Mittlere RPRn der Wiedererkennungsverfahren aus der Evaluation mit dem Fahndungsdatensatz. In der zweiten Spalte von rechts ist das KovIDent-Ergebnis Verfahren nach der Neusortierung mittels der paarweisen geodätischen Abständen (Referenz-Neusortierung) aufgeführt und in der rechten Spalte das KovIDent-Ergebnis nach der Neusortierung in der Einbettung. Die besten Ergebnisse sind fett hervorgehoben.

7.3.2 Evaluation mit dem Fahndungsdatensatz

Die Evaluation anhand des Fahndungsdatensatzes erfolgte analog zu der Evaluation mit dem Kameranetzwerk-Datensatz: Es wurde wieder je ein Tracklet aus der Galerie genommen und mit den restlichen verglichen. Wie auch beim Kameranetzwerk-Datensatz gibt es im Fahndungsdatensatz zu einigen Personen mehrere Tracklets. Die Evaluation erfolgte auch wieder unter *erschwerten* Bedingungen, so dass es zu einem Anfrage-Tracklet immer nur ein richtiges Galerie-Tracklet gab. Im Gegensatz zur vorherigen Evaluation wurden die Neusortierungen in dieser Evaluation auf dem gesamten Datensatz durchgeführt, um auch einen besseren Vergleich zwischen einer MSA auf den einzelnen Kovarianzdeskriptoren und KovIDent zu erhalten. Die Ergebnisse sind in der Tabelle 7.2 zusammengefasst, sowohl die mittleren RPRn des KovIDent-Verfahrens ohne Neusortierung (Ergebnis nach Schritt 4) als auch die mittleren RPRn der Referenz-Neusortierung und der Neusortierung in der Einbettung.

Die Neusortierung in der *Einbettung* erzielte in den oberen Ränge wieder bessere Ergebnisse als die Referenz-Neusortierung. Die Referenz-Neusortierung lieferte sogar deutlich schlechtere Ergebnisse als das

KovIDent-Verfahren im 3. Schritt. Ein Grund für die schlechteren Ergebnisse sind Ausreißer: Einige Einzelbilder zeigen keine oder eine falsche Person. Die Neusortierung in der Einbettung erzielte hingegen bessere Ergebnisse, da im Zielraum die Ausreißer besser berücksichtigt wurden.

7.4 Zusammenfassung

Durch die Kombination der erscheinungsbasierten Personenrepräsentation mittels Kovarianzdeskriptoren mit der unüberwachten Tracklet-Deskriptor-Bestimmung und der MaL basierten Neusortierung von Teilergebnissen wurde ein neuartiger, effizienter, bildsequenzbasierter Ansatz zur Überprüfung, ob mehrere Ganzkörpersequenzen von derselben Person stammen, erarbeitet. Der Ansatz übertrifft einen konventionellen unüberwachten Ansatz und kann zudem einfach für andere Merkmale angepasst werden. Der Tracklet-Deskriptor bestimmt sich, im Gegensatz zu den meisten mittelwertbasierten Ansätzen, aus mehreren Mittelwerten, wobei eine unüberwachte Strategie für die Bestimmung der Anzahl der Mittelwerte für ein Tracklet mittels Spectral Clustering Methoden verfolgt wird. Zudem ist der Ansatz nicht auf zeitlich zusammenhängende Bildsequenzen beschränkt: Es können auch Tracklet-Deskriptoren aus einzelnen Bildausschnitten von z.B. unterschiedlichen Quellen bestimmt werden. Die MaL basierte Neusortierung der Teilergebnissen hat bestätigt, dass durch die Annahme, dass die Kovarianzdeskriptoren auf einer Mannigfaltigkeit liegen, weitere Verbesserungen erzielt werden können (vgl. Kapitel 5). Die Evaluationen anhand zweier Datensätze für die kameraübergreifende Wiedererkennung und Personensuche in Bilddatensätzen, die sowohl verschiedene Ausprägungen der Personenerscheinung als auch mangelnde Bildqualität aufweisen, ergaben, dass die Ansätze jeweils eine Verbesserung erzielten.

8

KovIDent und tiefe künstliche faltende neuronale Netze (TKFNN)

Wegen der rasanten Entwicklung der Verarbeitungsgeschwindigkeit von Grafikkarten und anderen PC-Komponenten sowie der steigenden Anzahl zur Verfügung stehender Bilddatenbanken ist es heute möglich, sehr tiefe künstliche faltende neuronale Netze in akzeptabler Zeit zu trainieren (siehe z.B. [Den09, Kri12]). Aufgrund ihrer guten Ergebnisse in vielen Bereichen der Bildauswertung, wie z.B. in der erscheinungsbasierten Personenwiedererkennung, verzeichnet der Einsatz von TKFNN eine Zunahme (siehe [Xia16]). Bei der Wiedererkennung werden Bildausschnitte von Personen als Eingabe in die TKFNN und eine zugehörige Klasse, die die Identität der Person repräsentiert, als Ausgabe verwendet. Je mehr Datenpaare (Bild-Klassen-Paare) dabei zur Verfügung stehen, desto tiefer kann das TKFNN trainiert werden, was in der Regel zu einer höheren Wiedererkennungsrate führt. Mit der heute zur Verfügung stehenden Hardware können innerhalb weniger Stunden oder Tage sehr tiefe künstliche neuronale Netze auf sehr vielen Datenpaaren trainiert werden, die sehr gute Wiedererkennungsraten liefern. Die TKFNN müssen allerdings, um gute Ergebnisse erzielen zu können, auf einer ausreichend großen Trainingsdatenmenge trainiert werden.

In diesem Kapitel werden Möglichkeiten der Verknüpfung von handentworfenen Merkmalen, die den Kern dieser Dissertation bilden, und mittels TKFNN gelernten Merkmalen (TKFNN-Merkmale) präsentiert. Es wird eine Methode — im Folgenden als *Kov-TKFNN* bezeichnet — vorgestellt, die durch TKFNN eine verbesserte, mittelwertbasierte Personenrepräsentation auf Basis von Kovarianzdeskriptoren erzielt. Außerdem wird ein Ansatz — im Folgenden als *Fusion-TKFNN* bezeichnet — vorgestellt, der eine handentworfene Strategie mit TKFNN basierten Verfahren verknüpft [Sch17]. Das Ziel ist es, mittels *tiefer Fusion* von handentworfenen und gelernten Merkmalen eine höhere Wiedererkennungsrate als bei den einzelnen Ansätzen zu erlangen. Unter tiefer Fusion ist hier die Fusion der Merkmale während des Trainings eines TKFNN zu verstehen. Als handentworfene Merkmale bzw. Deskriptoren werden aufgrund der vorangegangenen Ergebnisse wieder Kovarianzdeskriptoren verwendet, wobei sich dieser Ansatz nicht auf diese beschränkt. Die beiden Ansätze haben das Potential, sowohl konventionelle Methoden als auch die populären TKFNN basierten Verfahren zu verbessern.

Das Ziel beider Verfahren ist es, ähnlich wie im vorherigen Kapitel, eine Person anhand von niedrig aufgelöstem Bildmaterial in Bilddaten unabhängig von der Bildquelle, dem Aufnahmeort und der Aufnahmezeit wiederzufinden. Beide Verfahren werden hinsichtlich einzelbildbasierter Personenwiedererkennung evaluiert. Einzelbildbasiert bedeutet, dass als Anfrage nur ein Bildausschnitt einer Person zu Verfügung steht und mit einzelnen Bildausschnitten aus einer Datenbank abgeglichen wird, die ein Einzelbild oder wenige Einzelbilder der gesuchten Person beinhaltet.

8.1 Kov-TKFNN

Die Grundidee des Kov-TKFNN ist es, ein TKFNN mit Bildausschnitten von Personen als Eingabe und vorberechneten Kovarianzdeskriptoren als Ausgabe zu trainieren, wobei die vorberechneten Kovarianzdeskriptoren den logarithmierten Mittelwerten entsprechen, die aus den Kovarianzdeskriptoren mehrerer Bildausschnitte einer Person berechnet werden. In den vorherigen beiden Kapiteln wurde gezeigt, dass eine mittelwertbasierte Repräsentation sehr gut mehrere Ansichten einer Personen repräsentieren

kann. Durch Kov-TKFNN sollen für Einzelbilder Kovarianzdeskriptoren bestimmt werden können, die näher an dem — a priori nicht bekannten — Mittelwert sind, der die Person am *besten* repräsentiert. Durch die Verwendung von logarithmierten Werten können reguläre TKFNN trainiert werden, da der gewöhnliche Logarithmusoperator log von Matrizen den Raum der Kovarianzdeskriptoren Sym_n^+ zu einer flachen riemannschen Mannigfaltigkeit reduziert [Ars07] (vgl. Kapitel 4). Eine flache riemannsche Mannigfaltigkeit ist lokal isometrisch zum euklidischen Raum, weshalb die logarithmierten Kovarianzdeskriptoren durch Vektoren repräsentiert werden können.

Training. Seien $\{\boldsymbol{I}_{\boldsymbol{R}_i}\}$, $i = 1, \ldots, n_p$ Bildausschnitte einer Person und $\bar{\boldsymbol{\Sigma}}$ der Mittelwert der aus den Bildausschnitten berechneten Kovarianzdeskriptoren $\{\boldsymbol{\Sigma}_{\boldsymbol{I}_{\boldsymbol{R}_i}}\}$, $i = 1, \ldots, n$. Das Training des Netzes erfolgt durch Einzelbild-Mittelwert-Paare, d.h. durch die paarweise Eingabe der Einzelbilder einer Person in die Eingangsschicht und dem jeweiligen zugehörigen Mittelwert der Kovarianzdeskriptoren in die Merkmalsebene. Das Konzept ist in Abbildung 8.1 illustriert. Dabei werden allerdings nicht die Mittelwerte der Kovarianzdeskriptoren direkt verwendet, sondern deren logarithmierte Mittelwerte: $\log(\bar{\boldsymbol{\Sigma}})$. In der Merkmalsebene werden diese durch Vektoren repräsentiert. Die Umrechnung erfolgt mit dem $\mathrm{vec}_{\mathrm{OD}}$-Operator, der aus der oberen Dreiecksmatrix einer $n \times n$-Kovarianzmatrix $\boldsymbol{\Sigma}$ einen $\frac{n\cdot(n+1)}{2}$-dimensionalen Vektor $\boldsymbol{c}$ bestimmt:

$$\boldsymbol{c} := \begin{pmatrix} \sigma_{1,1} \\ \sigma_{1,2} \\ \sigma_{2,2} \\ \sigma_{1,3} \\ \sigma_{2,3} \\ \vdots \\ \sigma_{n,n} \end{pmatrix} = \mathrm{vec}_{\mathrm{OD}} \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \ldots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \ldots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \ldots & \sigma_{n,n} \end{pmatrix} = \mathrm{vec}_{\mathrm{OD}}(\boldsymbol{\Sigma})\,. \quad (8.1)$$

Entsprechend wurde eine euklidische Zielfunkion für das Training verwendet.

Für das Training wurde der Datensatz *CUHK3*[1] [Li14] verwendet. Der Datensatz beinhaltet Bildausschnitte von 1360 Personen, die aus Bildse-

[1] `http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html`

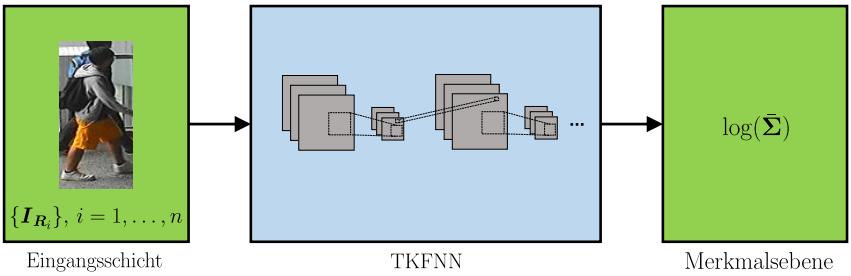


Abbildung 8.1: Konzept des Kov-TKFNN Ansatzes.

Abbildung 8.2: Beispielbildpaare 5 unterschiedlicher Personen aus dem CUHK3-Datensatz.

quenzen unterschiedlicher Kameras ausgeschnitten wurden. Die Bildausschnitte unterscheiden sich zwischen den Kameras deutlich hinsichtlich Aufnahmewinkel und Beleuchtung. Die Bildausschnitte wurden für das Training auf die Größe von 64 Pixel (Breite) × 80 Pixel (Höhe) skaliert. In Abbildung 8.2 sind Beispielbildausschnitte von 5 Personen dargestellt.

Online. Der Vergleich zweier Bildausschnitte $\boldsymbol{I}_{\boldsymbol{R}}$ und $\boldsymbol{J}_{\boldsymbol{R}}$ erfolgt anhand der Kosinus-Ähnlichkeit der zugehörigen logarithmierten Kovarianzdeskriptoren $\boldsymbol{c}_{\boldsymbol{I}} = \mathrm{vec}_{\mathrm{OD}}(\log(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{I}_{\boldsymbol{R}}}))$ und $\boldsymbol{c}_{\boldsymbol{J}} = \mathrm{vec}_{\mathrm{OD}}(\log(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{J}_{\boldsymbol{R}}}))$:

$$k(\boldsymbol{c}_{\boldsymbol{I}}, \boldsymbol{c}_{\boldsymbol{J}}) = \frac{\boldsymbol{c}_{\boldsymbol{I}}{}^{T} \cdot \boldsymbol{c}_{\boldsymbol{J}}}{||\boldsymbol{c}_{\boldsymbol{I}}||_2 \cdot ||\boldsymbol{c}_{\boldsymbol{J}}||_2} \tag{8.2}$$

Im Gegensatz zum KovIDent-Verfahren werden beim Kov-TKFNN in der Onlinephase die Mittelwerte durch das TKFNN bestimmt, das Bildausschnitte $\boldsymbol{I}_{\boldsymbol{R}}$ als Eingabe erhält und logarithmierte Kovarianzdeskriptoren $\log(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{I}_{\boldsymbol{R}}})$ ausgibt.

Schichten	Größe	Schritt-weite / Padding	Dimension der Ausgabe
Eingabe			$3 \times 64 \times 80$
Faltungsschicht 1	7×7	1 / 0	$40 \times 58 \times 74$
Gruppierungsschicht 1	2×2	2 / -	$40 \times 29 \times 37$
Faltungsschicht 2	5×5	1 / 0	$144 \times 25 \times 33$
Gruppierungsschicht 2	2×2	2 / -	$288 \times 13 \times 17$
Faltungsschicht 3	2×2	1 / 0	$32 \times 12 \times 16$
$\boldsymbol{s}_i$ (Merkmalsebene)			66

Tabelle 8.1: TKFNN-Architektur zur Berechnung der logarithmierten Kovarianzdeskriptoren. Die Dimension der Ausgabe ist wie folgt angegeben: Kanäle × Breite × Höhe.

Netzarchitektur. Die verwendete Netzarchitektur ist in der Tabelle 8.1 zusammengefasst. Die Eingangsschicht wurde für Bildausschnitte mit einer Größe von 64 Pixel (Breite) ×80 Pixel (Höhe) definiert. Zur Bestimmung der logarithmierten Kovarianzdeskriptoren werden 3 Faltungsschichten verwendet. Die Dimensionsreduktion erfolgt mittels Schritterhöhung zweier Gruppierungsschichten auf 2 (vgl. [Sze16]). Weitere Dimensionsreduktionen werden aufgrund der niedrigen Auflösung der Bildausschnitte nicht durchgeführt. Die Ausgabe sind 66-dimensionale Vektoren $\boldsymbol{s}_i$, die die oberen bzw. unteren Dreiecksmatrizen der logarithmierten Kovarianzdeskriptoren repräsentieren.

8.2 Fusion-TKFNN

In diesem Abschnitt wird der Fusionsansatz präsentiert, der im Folgenden als Fusion-TKFNN bezeichnet wird und kurz zusammengefasst wird. Der Ansatz wurde in [Sch17] veröffentlicht, worauf für eine ausführliche

Darstellung des Fusionsansatzes verwiesen wird. Fusion-TKFNN basiert auf dem Ziel, die Stärken von handentworfenen und gelernten Merkmalen (TKFNN-Merkmale) zu verknüpfen. Dafür werden die handentworfenen Merkmale, die nicht gelernt werden müssen, in den Trainingsprozess der TKFNN-Merkmale mit eingeschlossen, was beispielsweise einer Überanpassung beim Training entgegenwirkt. Außerdem ändert sich dadurch der *Fokus* des TKFNN von der allgemeinen Personenwiederkennung hin zum Lernen Fälle zu kompensieren, in denen die handentworfenen Merkmale schlechte Ergebnisse liefern. Die tiefe Fusion kann mittels drei unterschiedlicher Fusionsansätze durchgeführt werden, die in Abschnitt 8.2.2 behandelt werden. Davor werden im folgenden Abschnitt zunächst noch die verwendeten TKFNN-Merkmale vorgestellt.

8.2.1 TKFNN-Merkmale

Die Netzarchitektur in [Sch17] für die Berechnung der TKFNN-Merkmale besteht hauptsächlich aus *Inception*-Blöcken [Sze15], die aus mehreren Faltungsschichten aufgebaut sind, und ähnelt der Architektur in [Xia16] und den Netzen, die für Kov-TKFNN verwendet wurden.

Als Eingabebilder müssen die Bildausschnitte auf eine Größe von 64 Pixel (Breite) $\times$160 Pixel (Höhe) skaliert werden, was einer Eingangsdimension von 30720 entspricht. Für die Extraktion der Basismerkmale werden vier Faltungsschichten verwendet. Darauf folgen vier Gruppen mit jeweils zwei Inception-Blöcke: Einem regulären Inception-Block mit 3×3 Faltungsschichten anstatt 5×5 und einem Inception-Block zur Dimensionsreduktion der Merkmalskarte. Die Dimensionsreduktion erfolgt mittels Schritterhöhung der letzten Gruppierungs- und Faltungsschichten auf 2 [Sze16]. Weitere Dimensionsreduktionen werden aufgrund der niedrigen Auflösung der Bildausschnitte nicht durchgeführt.

Die TKFNN-Merkmale entsprechen dann der Ausgabe einer 256-dimensionalen voll vernetzten Schicht (*TKFNN-Merkmalsebene*). Die letzte Schicht des TKFNN kombiniert eine Softmax-Funktion mit einer Kostenfunktion, deren Dimension der Anzahl der Personen entspricht. Eine Übersicht über die Netzarchitektur ist in der Tabelle 8.2 gegeben.

Schichten	**Größe**	**Schritt-weite / Padding**	**Dimension der Ausgabe**	**IBK**
Eingabe			3 × 64 × 160	
Faltungsschichten 1-4	3 × 3	1 / 1	32 × 64 × 160	
Gruppierungsschicht	2 × 2	2 / -	32 × 32 × 80	
Inception-Block 1a			256 × 32 × 80	64
Inception-Block 1b		2 / 1	384 × 16 × 40	64
Inception-Block 2a			512 × 16 × 40	128
Inception-Block 2b		2 / 1	768 × 8 × 20	128
Inception-Block 3a			1024 × 8 × 20	256
Inception-Block 3b		1 / 1	1536 × 8 × 20	256
Inception-Block 4a			1024 × 8 × 20	256
Inception-Block 4b		2 / 1	1536 × 4 × 10	256
TKFNN-Merkmalsebene			256	

Tabelle 8.2: TKFNN-Architektur des Fusion-TKFNN. Die Dimension der Ausgabe ist wie folgt angegeben: Kanäle × Breite × Höhe. IBK geben die Kanäle (Dimension) der einzelnen Pfade innerhalb eines Inception-Blockes an.

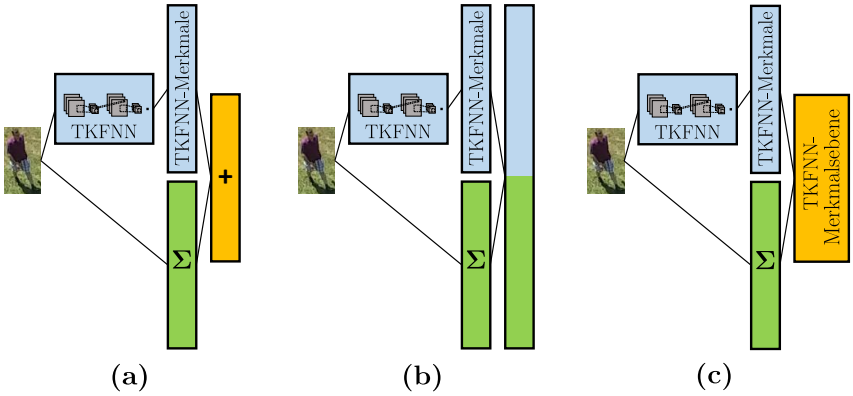


Abbildung 8.3: Unterschiedliche Ansätze für die Tiefe Fusion: tiefe Addition **(a)**, tiefe Konkatenation **(b)** und tiefe vv-Fusion **(c)**.

8.2.2 Tiefe Fusion

Für Fusion-TKFNN werden drei unterschiedliche Ansätze, die in Abbildung 8.3 dargestellt sind, zur Einbindung von handentworfenen Merkmalen in den Trainingsprozess der TKFNN-Merkmale vorgeschlagen (vgl. [Sch17]): *tiefe Addition*, *tiefe Konkatenation* und *tiefe voll vernetzte Fusion* (tiefe vv-Fusion).

Tiefe Addition. Die tiefe Addition, die in Abbildung 8.3a dargestellt ist, ist ein einfacher Fusionsansatz, bei der die handentworfenen Merkmale mit den TKFNN-Merkmalen elementweise addiert werden. Der Nachteil dabei ist, dass die Dimension der handentworfenen Merkmale mit der Dimension der gelernten Merkmale übereinstimmen muss. Die tiefe Addition ähnelt den Architekturen der Residuum Netze [He16], bei denen abkürzende Verbindungen eingefügt werden. Angelehnt an diese Architektur werden bei der tiefen Addition logarithmierte Kovarianzdeskriptoren als Abkürzungen eingefügt, die mit dem Hauptpfad dann elementweise verknüpft werden.

Tiefe Konkatenation. Dieser Fusionsansatz verknüpft die handentworfenen Merkmale mit den TKFNN-Merkmalen durch eine Konkatenati-

onsschicht, was einer konventionellen Konkatenation von Merkmalen entspricht, allerdings mit dem Unterschied, dass bei der tiefen Konkatenation die handentworfenen Merkmale schon während des Trainings berücksichtigt werden (siehe Abbildung 8.3b). Bei der Konkatenation müssen die Dimensionen der unterschiedlichen Merkmale nicht übereinstimmen und das resultierende Merkmal hat eine deutlich größere Dimension.

Tiefe vv-Fusion. Wie in [Wu16] vorgeschlagen und in Abbildung 8.3c dargestellt, wird bei der tiefen vv-Fusion eine voll vernetzte Schicht verwendet, um die handentworfenen Merkmale im Trainingsprozess zu integrieren. In Vergleich zu den beiden vorherigen tiefen Fusionsansätzen erhöht dieser Ansatz die Parameteranzahl in der Netzarchitektur. Um die Parameteranzahl gering zu halten, wurde eine zusätzliche voll vernetzte Schicht gewählt, welche die Dimension des resultierenden Merkmals auf 256 setzt.

8.3 Verfahrensevaluation

Kov-TKFNN wurde auf drei umfangreichen, öffentlichen Personenwiedererkennungsdatensätzen für einzelbildbasierte Verfahren evaluiert: *VIPeR* [Gra07], *CUHK1* [Li12b] und *CAVIAR4REID* [Che11].

Der Fusion-TKFNN wurde auf einem *internen* Luft-Boden-Bilddatensatz evaluiert, der explizit zum Testen von einzelbildbasierten Wiedererkennungsverfahren akquiriert wurde [Sch17]. Der Luft-Boden-Bilddatensatz ist insbesondere für die Evaluation von Verfahren für die Wiedererkennung von Personen zwischen Luft- und Bodenbildern erstellt worden.

VIPeR. Der Datensatz VIPeR[2] [Gra07] besteht aus 632 Personenbildpaaren, wobei sich die einzelnen Bildpaare deutlich hinsichtlich Aufnahmewinkel und Beleuchtung unterscheiden. Die Bilder eines Bildpaars stammen von unterschiedlichen Kameras, die im Originaldatensatz auf eine feste Größe skaliert wurden: 48 Pixel (Breite) × 128 Pixel (Höhe). In Abbildung 8.4 sind 11 Beispiele von den 632 Personenbildpaaren dargestellt.

[2]Der VIPeR-Datensatz ist unter folgendem Link zum Download verfügbar: `https://vision.soe.ucsc.edu/node/178`

Abbildung 8.4: Beispielbildpaare aus dem VIPeR-Datensatz. Eine Spalte entspricht einem von 632 Personenbildpaaren.

CUHK1. Der Datensatz CUHK1[3] [Li12b] beinhaltet Bildausschnitte von 971 Personen, die aus Sequenzen zweier unterschiedlicher Kameras ausgeschnitten wurden. Zu jeder Person gibt es vier Bildausschnitte, wovon jeweils zwei Bildausschnitte aus Sequenzen einer Kamera stammen. Die Bildausschnitte von der einen Kamera unterscheiden sich ähnlich wie beim VIPeR-Datensatz deutlich zu den Bildausschnitten der zweiten Kamera hinsichtlich Aufnahmewinkel und Beleuchtung. Die Originalauflösung beträgt 64 Pixel (Breite) × 160 Pixel (Höhe). In Abbildung 8.5 sind Beispielbildausschnitte von 11 Personen dargestellt.

CAVIAR4REID. Der Datensatz CAVIAR4REID[4] [Che11] beinhaltet Bildausschnitte von 72 unterschiedlichen Personen, wovon 50 Personen nur in einer Kamera und 22 in zwei Kameras zu sehen sind. Welcher Bildausschnitt von welcher Kamera stammt ist im Gegensatz zum VIPeR- und CUHK1-Datensatz nicht angegeben. Im Gegensatz zum VIPeR-Datensatz, der je Person nur zwei Bilder im Datensatz hat, gibt es im CAVIAR4REID-Datensatz mehrere Bilder zu einer Person. Insgesamt gibt es 1220 Bildausschnitte zu den Personen, die aus 26 Bildsequenzen von Innenaufnahmen ausgestanzt wurden. Die Anzahl der Bilder pro

[3] `http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html`

[4] Der CAVIAR4REID-Datensatz ist unter folgendem Link zum Download verfügbar: `http://www.lorisbazzani.info/caviar4reid.html`

Abbildung 8.5: Beispielbildausschnitte von 11 Personen aus dem CUHK1-Datensatz. Die oberen und unteren beiden Reihen zeigen spaltenweise jeweils zwei Bilder einer Person aus einer Kamera.

Abbildung 8.6: Beispielbildausschnitte von 11 Personen aus dem CAVIAR4REID-Datensatz.

Person unterscheidet sich von Person zu Person. Die Auflösungen der Bildausschnitte variieren zwischen einer Auflösung von 17 Pixel (Breite) × 39 Pixel (Höhe) und 72 Pixel (Breite) × 144 Pixel (Höhe). Abbildung 8.6 zeigt Beispielbildausschnitte von 11 unterschiedlichen Personen.

Luft-Boden-Bilddatensatz. Der Luft-Boden-Bilddatensatz umfasst Bilddaten aus der Vogelperspektive (teilweise Nadir-Sicht) und Schrägsichten von Kameras, die auf Laternen o.ä. befestigt waren. Er ist auf einen *Luft*-Teildatensatz und vier *Boden*-Teildatensätze aufgeteilt, so dass neben einer Boden-zu-Luft- und Luft-zu-Boden-Evaluation auch Boden-zu-Boden-Evaluationen durchgeführt werden können. Die Luftaufnahmen wurden mit Hilfe eines Quadrokopters aufgezeichnet, so dass sich die Blickwinkel aus der Vogelperspektive unterscheiden. Im Gegensatz zu den Wiedererkennungsdatensätzen in Kapitel 7 existieren in diesem Datensatz — und auch bei den öffentlichen Datensätzen — keine Bildsequenzen von Personen, wodurch er nur für die Evaluation eines Einzelbildvergleichs relevant ist. Insgesamt umfasst der Luft-Boden-Bilddatensatz 1217 Anfrage- und 4244 Galerie-Bildausschnitte von 14 unterschiedlichen Personen, die aus Bildsequenzen von vier unterschiedlichen Kameras ausgestanzt wurden. Für die Evaluation wurde der Datensatz *ausbalanciert*: Es wurden maximal 10 Bildausschnitte je Person für die Anfrage und maximal 100 Bildausschnitte je Person für die Galerie verwendet. Die Auflösungen der Personen sind aufgrund der unterschiedlichen Entfernungen und Blickwinkel stark verschieden. Die Luftbilder zeigen beispielsweise Personen, die deutlich geringer aufgelöst sind. Die Auflösung einer Person liegt zwischen ca. 40 und 200 Pixel in der Breite und zwischen ca. 150 und 600 Pixel in der

Abbildung 8.7: Bilder aus dem Luft-Boden-Bilddatensatz von sechs unterschiedlichen Personen. Je Reihe ist eine Person zu sehen. In der Übersicht sind Beispielbilder von drei verschiedenen Kameras dargestellt. Die farbigen (Teil-)Rahmen kennzeichnen die Kameras: blau (Vogelperspektive), grün(Bodenkamera Laterne) und orange (Bodenkamera Empore).

Höhe. Abbildung 8.7 zeigt beispielhaft sechs Personen des Luft-Boden-Bilddatensatzes aus drei unterschiedlichen Blickwinkeln.

Evaluations-Metrik. Als Evaluations-Metrik wird der *mittlere positive Vorhersagewert* (MPV) verwendet. Der MPV ist der Mittelwert der in Abschnitt 6.3.1 beschriebenen positiven Vorhersagewerte von Q Anfragen:

$$\mathrm{MPV} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{PV}(q)\,, \tag{8.3}$$

wobei $\mathrm{PV}(q)$ dem positiven Vorhersagewert der Anfrage q entspricht. Bei der Kov-TKFNN Evaluation werden für eine genauere Betrachtung des resultierenden Rankings zudem die RPRn berechnet.

8.3.1 Kov-TKFNN Evaluation

Im Rahmen der Evaluation des Kov-TKFNN wurden die Merkmalsvektoren für die Körperteildetektion um Farbinformationen ergänzt. Es wurden somit 11-dimensionale Kovarianzdeskriptoren betrachtet:

$$\boldsymbol{f}_{(x,y)} = \begin{pmatrix} X(x,y) \\ Y(x,y) \\ L_{Lab}(x,y) \\ a_{Lab}(x,y) \\ b_{Lab}(x,y) \\ \left|\frac{\delta \boldsymbol{I}(x,y)}{\delta x}\right| \\ \left|\frac{\delta \boldsymbol{I}(x,y)}{\delta y}\right| \\ \left|\frac{\delta \boldsymbol{I}(x,y)}{\delta x 2}\right| \\ \left|\frac{\delta \boldsymbol{I}(x,y)}{\delta y^2}\right| \\ \sqrt{(\boldsymbol{I}_x(x,y)^2 + \boldsymbol{I}_y(x,y)^2)} \\ \tan^{-1}\left(\left|\frac{\boldsymbol{I}_y(x,y)}{\boldsymbol{I}_x(x,y)}\right|\right) \end{pmatrix}, \tag{8.4}$$

mit $X(x,y) = x$ und $Y(x,y) = y$. Die Gradienten werden auf den Intensitätsbildern berechnet.

Die Evaluation des Kov-TKFNN erfolgte anhand der folgenden drei typischen Schritte einer Personenwiedererkennung für Suchsysteme:

1. Bestimmung der Deskriptoren (logarithmierte Mittelwerte der Kovarianzdeskriptoren berechnen bzw. mittels TKFNN bestimmen),

2. Ähnlichkeit der Deskriptoren berechnen (anhand der Kosinus-Ähnlichkeit) und

3. Erstellung einer sortierten Rangliste anhand der Ähnlichkeiten.

Datensatz	MPV	Rang 1	Rang 5	Rang 10	Rang 20
VIPeR	0,072	0,029	0,099	0,127	0,187
CUHK1	0,071	0,031	0,097	0,133	0,185
CAVIAR4REID	0,278	0,178	0,348	0,457	0,639

Tabelle 8.3: Wiedererkennungsergebnis bei Verwendung von **berechneten** Kovarianzdeskriptoren: MPV und RPRn für die Ränge 1, 5, 10 und 20.

Datensatz	MPV	Rang 1	Rang 5	Rang 10	Rang 20
VIPeR	0,105	0,060	0,114	0,190	0,275
CUHK1	0,048	0,016	0,059	0,092	0,158
CAVIAR4REID	0,279	0,164	0,354	0,543	0,796

Tabelle 8.4: Wiedererkennungsergebnis bei Verwendung von **gelernten** Kovarianzdeskriptoren: MPV und RPRn für die Ränge 1, 5, 10 und 20. Bei dieser Evaluation wurde das TKFNN mit logarithmierten — **nicht** gemittelten — Kovarianzdeskriptoren der Einzelbilder trainiert.

Die Evaluation wurde auf den Datensätzen VIPeR, CUHK1 und CAVIAR4REID durchgeführt, wobei die 11-dimensionalen Kovarianzdeskriptoren durch das in Abschnitt 8.1 beschriebene TKFNN für diese Datensätze bestimmt wurden. Das TKFNN wurde dafür auf Einzelbild-Mittelwert-Paare trainiert, die aus dem CUHK3-Datensatz berechnet wurden (vgl. Abschnitt 8.1). Aufgrund des Anwendungsschwerpunkts dieser Arbeit wurden alle Bilder der Datensätze auf eine (niedrige) Auflösung von 64 Pixel (Breite) × 80 Pixel (Höhe) skaliert. Der MPV wurde mittels der Kosinus-Ähnlichkeit (Gleichung (8.2)) für jeweils den gesamten Datensatz ermittelt. Zu Vergleichsgründen wurde diese Evaluation wiederholt mit *berechneten* (Tabelle 8.3) und *gelernten* — nicht gemittelten – Kovarianzdeskriptoren durchgeführt (Tabelle 8.4). Bei allen drei Evaluationen wurde die Kosinus-Ähnlichkeit auf den oberen bzw. unteren Dreiecksmatrizen der logarithmierten Kovarianzdeskriptoren berechnet. Die Ergebnisse des Kov-TKFNN ist in der Tabelle 8.5 zusammengefasst.

Datensatz	MPV	Rang 1	Rang 5	Rang 10	Rang 20
VIPeR	0,104	0,041	0,152	0,206	0,310
CUHK1	0,081	0,036	0,103	0,158	0,234
CAVIAR4REID	0,287	0,167	0,384	0,565	0,777

Tabelle 8.5: Kov-TKFNN-Wiedererkennungsergebnis: MPV und RPRn für die Ränge 1, 5, 10 und 20. Hierbei wurde ein TKFNN mit **gemittelten** Kovarianzdeskriptoren (logarithmiert) trainiert.

Auf allen drei Datensätzen erzielten die mittels Kov-TKFNN gelernten Kovarianzdeskriptoren bessere MPV als die berechneten Deskriptoren (vgl. die Tabellen 8.3 und 8.5). Auch die RPRn konnten — bis auf die RPR des ersten Rangs beim CAVIAR4REID-Datensatz — gesteigert werden. Einen direkten Zusammenhang mit der Anzahl der Trainingsbilder bzw. Deskriptoren pro Person resultiert nicht aus diesen Ergebnissen. Der Vergleich der Ergebnisse in den Tabellen 8.4 und 8.5 bekräftigt zudem die Idee von Kov-TKFNN, dass durch die *gelernte* Repräsentation auf Basis eines Trainings mit *gemittelten* Kovarianzdeskriptoren eine verbesserte Personenrepräsentation erzielt wird. Dies liegt vermutlich daran, dass die daraus resultierenden Kovarianzdeskriptoren näher an den Mittelwerten liegen, welche die Person am *besten* repräsentieren. Dieser mittelwertbasierte Repräsentationsansatz kann eventuell in einer aufbauenden Folgearbeit weiter untersucht und verbessert werden.

8.3.2 Fusion-TKFNN Evaluation

In diesem Abschnitt werden zunächst die verwendeten Kovarianzdeskriptoren und TKFNN-Merkmale beschrieben. Die Bilder des betrachteten Luft-Boden-Bilddatensatzes wurden auf eine Auflösung von 64 Pixel (Breite) × 160 Pixel (Höhe) skaliert. Aufgrund des tieferen Netzes wurde eine größere Skalierung als bei der Evaluation des Kov-TKFNN gewählt. Außerdem werden die Evaluationsergebnisse aus [Sch17] zusammengefasst.

Kovarianzdeskriptoren. Die Kovarianzdeskriptoren, die im Rahmen dieser und den Evaluationen auf den öffentlichen Datensätzen verwendet wurden, basieren auf sechs Farbräumen. Diese Kovarianzdeskriptoren werden im Folgenden als *Multi-Farbraum-Deskriptoren* (MFD) bezeichnet. Der Merkmalsvektor $\boldsymbol{f}_{(x,y)}$ war für die Evaluationen wie folgt definiert:

$$\begin{aligned}\boldsymbol{f}_{(x,y)} = (&X(x,y),\, Y(x,y),\\ &R_{RGB}(x,y),\, G_{RGB}(x,y),\, B_{RGB}(x,y),\\ &H_{HSV}(x,y),\, S_{HSV}(x,y),\, V_{HSV}(x,y),\\ &L_{Lab}(x,y),\, a_{Lab}(x,y),\, b_{Lab}(x,y),\\ &Y_{YUV}(x,y),\, U_{YUV}(x,y),\, V_{YUV}(x,y),\\ &X_{XYZ}(x,y),\, Y_{XYZ}(x,y),\, Z_{XYZ}(x,y),\\ &Y_{YCrCb}(x,y),\, Cr_{YCrCb}(x,y),\, Cb_{YCrCb}(x,y))^T\,, \end{aligned} \tag{8.5}$$

mit $X(x,y) = x$ und $Y(x,y) = y$.

Im Gegensatz zum trackletbasierten Ansatz KovIDent werden aufgrund der stark unterschiedlichen Perspektiven — die insbesondere zwischen den Luft- und Bodenbildern vorhanden sind — neben den $y-$ auch die $x-$Pixelkoordinaten betrachtet. Außerdem wurden im Rahmen dieser Evaluation Kovarianzdeskriptoren verwendet, die sich aus mehreren Farbräumen berechnen. In [Mat16] wurde gezeigt, dass Kovarianzdeskriptoren, die sich aus verschiedenen Farbräumen berechnen, komplementäre Eigenschaften daraus gewinnen können. Zudem wurde für die Merkmalsauswahl diesbezüglich eine eigene Evaluation auf dem Luft-Boden-Bilddatensatz durchgeführt. Neben den Deskriptoren, die alle Farbräume berücksichtigen, wurden auch Deskriptoren evaluiert, die jeweils nur auf einem Farbraum basieren.

Das für die Bestimmung des MPV relevante Ranking im Rahmen der Evaluation der unterschiedlichen Kovarianzdeskriptoren wurde anhand der geodätischen Abstände ermittelt, die mit der Gleichung (4.8) berechnet wurden. Aus Vergleichsgründen wurden auch Evaluationen auf Basis der in Abschnitt 4.1.5 erwähnten log-euklidischen Metrik durchgeführt, die geringfügig schlechtere Ergebnisse lieferten. Für Details dazu wird auf die Veröffentlichung [Sch17] verwiesen. In dieser Veröffentlichung wurde zudem untersucht, ob die Erweiterung der farbraumbasierten Deskriptoren um

Kovarianz-deskriptor	**Ges. LBB**	**Luft (L)**	**Boden (B)**	**B-B**	**L-B**	**B-L**
HSV	0,207	0,289	0,388	0,160	0,171	0,182
Lab	0,301	0,314	0,452	0,276	0,260	0,265
RGB	**0,307**	0,325	0,476	**0,278**	0,260	**0,267**
YUV	**0,307**	0,325	0,478	0,277	0,260	0,266
XYZ	0,302	0,326	0,474	0,272	0,258	0,257
YCrCb	**0,307**	0,325	0,476	**0,278**	0,260	**0,267**
fusioniert (MFD)	0,296	**0,422**	**0,582**	0,217	**0,279**	0,232

Tabelle 8.6: MPV-Ergebnisse für die unterschiedlichen Kovarianzdeskriptoren auf dem Luft-Boden-Bilddatensatz (LBB). Wie die MFD berücksichtigen auch die Kovarianzdeskriptoren, die auf den einzelnen Farbräumen basieren, die $x-$ und $y-$Pixelkoordinaten. Für jeden Teildatensatz sind die besten Ergebnisse fett hervorgehoben.

Gradienteninformationen die Wiedererkennungsgenauigkeit weiter steigern kann, mit dem Ergebnis, dass keine Verbesserung erreicht wurde und folglich hier nicht weiter betrachtet wird.

Neben der Evaluation auf dem gesamten Datensatz wurden auch Evaluationen auf einzelnen Teildatensätze durchgeführt. Dazu wurde der gesamte Datensatz in die Teildatensätze *Luft* (**L**), *Boden* (**B**), *Boden-zu-Boden* (**B-B**), *Luft-zu-Boden* (**L-B**) und *Boden-zu-Luft* (**B-L**) unterteilt. Der MPV beim Teildatensatz **B** entspricht dem mittleren MPV von vier separaten Evaluationen, die jeweils innerhalb der Bilder einer *Bodenkamera* durchgeführt wurden. Bei den Teildatensätzen **B-B**, **L-B** und **B-L** wurde ausschließlich die kameraübergreifende Wiedererkennung evaluiert, d.h. Anfrage- und Galeriebild stammten immer von zwei unterschiedlichen Kameras.

Die verschiedenen farbraumbasierten Kovarianzdeskriptoren und das Evaluationsergebnis aus der Veröffentlichung [Sch17] ist in der Tabelle 8.6 zusammengefasst.

Die besten MPV erzielten die RGB-, YUV- und YCrCb-basierten Kovarianzdeskriptoren, die allerdings nur minimal besser als die der anderen sind (bis auf HSV), was aufgrund der einfachen Farbraumtransformationen zu

Merkmal	Gesamter LBB	Luft (L)	Boden (B)	B-B	L-B	B-L
TKFNN	**0,436**	0,372	**0,697**	**0,434**	**0,314**	**0,318**
MFD	0,296	**0,422**	0,582	0,217	0,279	0,232

Tabelle 8.7: MPV-Ergebnis für die TKFNN-Merkmale auf dem Luft-Boden-Bilddatensatz (LBB). Zum Vergleich ist in der untersten Zeile nochmal das MPV-Ergebnis der MFD aus Tabelle 8.6 aufgeführt.

erwarten war. Die RGB- und YCrCb-basierten Deskriptoren, die in allen Teildatensätzen gleich abschnitten, erzielten, wie auch die MFD, in drei Teildatensätzen den besten MPV. Die MFD ergaben zwar einen minimal geringeren durchschnittlichen MPV über den gesamten Testdatensatz, aber dafür eine deutlich höhere Wiedererkennungsgenauigkeit innerhalb der einzelnen Kameras, was vermutlich an den komplementären Eigenschaften der MFD liegt [Mat16]. In der folgenden Evaluation des Fusion-TKFNN wurden deshalb die MFD verwendet.

TKFNN-Merkmale. Zur Bestimmung der TKFNN-Merkmale wurde ein TKFNN auf dem *Market-1501*-Trainingsdatensatz[5] [Zhe15] trainiert. Der Trainings-Datensatz besteht aus 12.936 Bildausschnitten von 750 unterschiedlichen Personen. Die letzte Schicht des TKFNN kombiniert wieder eine Softmax-Funktion mit einer Kostenfunktion, deren Dimension gleich der Anzahl der Personen ist. Damit entspricht die Personenwiedererkennungsaufgabe der Aufgabe einer Klassifikation von Personenidentitäten. Das Evaluationsergebnis für die TKFNN-Merkmale aus der Veröffentlichung [Sch17] ist in der Tabelle 8.7 dargestellt.

Wie zu erwarten schneiden die TKFNN-Merkmale, im Vergleich zu den MFD, bei den höher aufgelösten Personen aus dem Teildatensatz **B** besser ab. Außer bei einem Teildatensatz konnten mit den gelernten Merkmalen bessere MPV erzielt werden. Bei der kameraübergreifenden Evaluation auf dem Datensatz **B-B** erzielten die gelernten Merkmale sogar ein deutlich besseres Ergebnis. Mit abnehmender Auflösung wurden die MFD gegenüber

[5] `http://www.liangzheng.org/Project/project_reid.html`

den TKFNN-Merkmalen *stärker*. Bei den Luftbildern haben die MFD 5 Prozentpunkte besser abgeschnitten.

Fusion-TKFNN Evaluation mit dem Luft-Boden-Bilddatensatz

In diesem Abschnitt werden die Ergebnisse aus der Evaluation mit dem Luft-Boden-Bilddatensatz zusammengefasst, die ausführlich in [Sch17] vorgestellt sind. Neben den Untersuchungen der drei tiefen Fusionsansätze, die in Abschnitt 8.2.2 vorgestellt wurden, sind zu Vergleichsgründen auch drei *konventionelle Fusionen* der Kovarianzdeskriptoren mit den TKFNN-Merkmalen durchgeführt worden. Die Evaluation erfolgt wieder anhand der Suchaufgabe.

Konventionelle Merkmalsfusion. Für den Vergleich der tiefen Fusion mit herkömmlichen Fusionsansätzen wurden folgende drei konventionelle Fusionsmethoden betrachtet:

- *merkmalsbasierte*,
- *ähnlichkeitsbasierte* und
- *entscheidungsbasierte* Fusion.

Bei der merkmalsbasierten Fusion werden die Merkmale nach der Merkmalsextraktion, also direkt nach Stufe 1, konkateniert. Dieser Ansatz wird auch als *frühe Fusion* bezeichnet und erfordert eine Metrik in der zweiten Stufe, die für beide Merkmale verwendet werden kann. Falls unterschiedliche Metriken gewünscht sind oder verwendet werden müssen, kann stattdessen ein ähnlichkeitsbasierter Fusionsansatz gewählt werden, der die Ähnlichkeiten bzw. die Abstände zwischen den Merkmalen fusioniert. Dies entspricht einer Fusion nach der zweiten Stufe und wird auch als *späte Fusion* bezeichnet. Bei diesem Fusionsansatz sollten die Wertebereiche der Abstände in einem ähnlichen Bereich liegen oder normalisiert werden, da andernfalls ein Merkmal dominieren würde. Der dritte konventionelle Fusionsansatz, die *entscheidungsbasierte Fusion*, ist die Fusion der Ranglisten durch beispielsweise der einfachen Bestimmung des mittleren Rangs.

Fusionsansatz	Ges. LBB	Luft (L)	Boden (B)	B-B	L-B	B-L
Merkmal	0,445	0,406	0,706	**0,435**	0,341	**0,329**
Ähnlichkeit	0,315	0,432	0,611	0,237	0,293	0,242
Ähnlichkeit Einheitslänge	0,435	**0,446**	**0,725**	0,404	**0,356**	0,315
Ähnlichkeit Min-Max	**0,448**	0,430	**0,725**	0,433	0,345	0,327
Ähnlichkeit Sigmoidfunktion	0,438	0,436	0,705	0,424	0,334	0,318
Entscheidung	0,391	0,426	0,680	0,344	0,333	0,295
Tiefe Addition	0,458	0,430	0,714	0,437	0,355	0,377
Tiefe Konkaten.	**0,478**	0,452	**0,725**	**0,466**	0,358	**0,390**
Tiefe vv-Fusion	0,447	**0,456**	0,665	0,422	**0,383**	0,369

Tabelle 8.8: MPV-Ergebnisse der konventionellen Fusion — merkmals- (Merkmal), ähnlichkeits- (Ähnlichkeit) und entscheidungsbasiert (Entscheidung) — sowie die Ergebnisse der tiefen Fusion von MFD und TKFNN-Merkmalen.

Bei der konventionellen Fusion der MFD mit den TKFNN-Merkmalen wurden die oberen Dreiecksmatrizen der logarithmierten Kovarianzdeskriptoren verwendet. Die Bestimmung der Ranglisten erfolgte mittels der Kosinus-Ähnlichkeit, die ein besseres Ergebnis lieferte als bei Verwendung des euklidischen Abstands [Sch17]. Die ähnlichkeitsbasierte Fusion wurde viermal durchgeführt, einmal ohne Normalisierung der Ähnlichkeitswerte und dreimal mit jeweils einer anderen Normalisierung: Normalisierung auf Einheitslänge, Min-Max-Normalisierung und Normalisierung mittels Sigmoidfunktion. Die Fusionsergebnisse der konventionellen Ansätze sind in der Tabelle 8.8 zusammengefasst.

Das schlechte Abschneiden der entscheidungsbasierten Fusion war zu erwarten. Sie entspricht zwar der ähnlichkeitsbasierten Fusion, es wird allerdings auf einer gröberen Ebene fusioniert: Es werden nur die Ränge und keine Ähnlichkeiten bzw. Abstände betrachtet. Die merkmals- und ähnlichkeitsbasierte Fusion haben ein besseres und zueinander ähnliches Ergebnis

erreicht, wobei die Min-Max-Normalisierung für das beste Ergebnis des ähnlichkeitsbasierten Ansatzes ausschlaggebend war.

Tiefe Fusion. Die Evaluation der in Abschnitt 8.2.2 vorgestellten tiefen Fusionsansätze *tiefe Addition*, *tiefe Konkatenation* und *tiefe vv-Fusion* erfolgte mittels der in Abschnitt 8.2.1 beschriebenen Architektur. Bei der *tiefen Addition* musste lediglich die Dimension der TKFNN-Merkmalsebene (letzte voll vernetzte Schicht) auf 210 angepasst werden, was der Dimension der oberen Dreiecksmatrizen der Kovarianzdeskriptoren entsprach. Die Ergebnisse der unterschiedlichen Ansätze für die tiefe Fusion sind ebenfalls in der Tabelle 8.8 zusammengefasst.

Die *tiefe Konkatenation* übertrifft alle konventionellen Fusionsverfahren hinsichtlich MPV und erzielt auch eine Verbesserung aller MPV der einzelnen nicht-fusionierten Ansätze (vgl. Tabelle 8.7). In der Kategorie der Luftbilder konnten die Schwächen der TKFNN-Merkmale und bei den Bodenbildern die Schwächen der Kovarianzdeskriptoren kompensiert werden. Auch bei der kameraübergreifenden Personenwiedererkennung konnte eine höhere MPV als bei den einzelnen nicht-fusionierten Ansätzen erreicht werden. Hinsichtlich Vergleich der drei tiefen Fusionsansätze erreichte die *tiefe Konkatenation* die beste Leistung. Der Grund dafür ist wahrscheinlich die erhöhte Dimensionalität der resultierende Merkmale bei der *tiefen Konkatenation*, die $466 (= 210 + 256)$ ist (vgl. [Sch17]).

8.4 Zusammenfassung

In diesem Kapitel wurden zwei lernbasierte Strategien für die erscheinungsbasierte Personenwiedererkennung vorgestellt. Die erste verfolgt einen mittelwertbasierten Ansatz, der aus unterschiedlichen Ansichten einer Person einen mittelwertbasierten Kovarianzdeskriptor lernt. Die Ergebnisse zeigen, dass die gelernten Kovarianzdeskriptoren diskriminativer sind bzw. *besser* die Personen repräsentieren als die entsprechenden berechneten Kovarianzdeskriptoren. Die zweite Strategie — die tiefe Fusion von Kovarianzdeskriptoren mit TKFNN-Merkmalen — verknüpft vorteilhaft zwei unterschiedliche Merkmalsarten, was durch die Ergebnisse einer Evaluation auf einem Luft-Boden-Bilddatensatz bekräftigt wird. Die Ergebnisse

motivieren weitere Untersuchungen in diese Richtung, die sowohl für die erscheinungsbasierte Personenwiedererkennung als auch für andere Bildauswerteaufgaben interessant sind.

9

Zusammenfassung und Ausblick

9.1 Zusammenfassung

Im Rahmen dieser Arbeit wurde ein Bildauswerte-Rahmenwerk für niedrig aufgelöste Bilder erarbeitet, das mehrere Personenrepräsentationsansätze und Methoden für vielfältige Bildauswerteaufgaben zur Verfügung stellt. Sowohl mittels der Betrachtung der Kovarianzdeskriptor-Untermannigfaltigkeiten als auch durch die neuartigen mittelwertbasierten Strategien konnten in drei Anwendungen — Personendetektion, -tracking und -wiedererkennung — Verbesserungen erzielt werden.

Im Rahmen der Personendetektion wurden drei verschiedene Repräsentationsarten miteinander verglichen: Körperteilklassifikation mit HOGs, Kovarianzdeskriptoren und Kovarianzdeskriptor-Untermannigfaltigkeiten. Der Ansatz auf Basis der Kovarianzdeskriptor-Untermannigfaltigkeiten erzielte im Gegensatz zu dem Basisverfahren (HOG-Ansatz) ein deutlich besseres und im Vergleich zur kovarianzdeskriptorbasierten Repräsentation ein geringfügig besseres Klassifikationsergebnis, was jedoch die Annahme der Untermannigfaltigkeiten im Sym_n^+ stärkt. Die Kovarianzdeskriptor-Untermannigfaltigkeiten werden gemeinsam mit einem überwachten Laplacian Eigenmaps Algorithmus gelernt, so dass in der resultierenden niedrigdimensionalen Gesamtrepräsentation die

Kovarianzdeskriptor-Untermannigfaltigkeiten verschiedener Körperteilklassen möglichst diskriminativ zueinander sind, was schließlich zu einer verbesserten Repräsentation für Klassifikationsaufgaben führen kann.

Außerdem wurde ein Kovarianz-Trackingverfahren durch statistische Erweiterungen robuster gestaltet. Im Vergleich zu bestehenden Kovarianz-Trackingverfahren wird für das Personenmodell, das sich über die Zeit der sich veränderten Erscheinung der Person anpasst, eine neue Aktualisierungsstrategie vorgeschlagen, die anhand statistischer Eigenschaften im Sym_n^+ einzelne Kovarianzdeskriptoren in diesem Schritt ausschließt. Durch diese Erweiterungen konnten bessere Ergebnisse erzielt werden, was auch den *robusten* mittelwertbasierten Tracklet-Deskriptor-Ansatz für die Personenwiederkennung motivierte.

Der bildsequenzbasierte Ansatz für die erscheinungsbasierte Personenwiedererkennung resultiert aus den Beobachtungen und Ergebnissen der Arbeiten zu Personendetektion und -tracking. Für die bildsequenzbasierte Methode wurde eine neuartige unüberwachte Strategie zur Bestimmung von Tracklet-Deskriptoren für die Repräsentation einzelner Personen-Tracklets vorgeschlagen. Es wird ein mittelwertbasierter Ansatz verfolgt, der mittels Spectral Clustering mehrere Mittelwerte für ein Tracklet bestimmt, wobei auch die Anzahl der Mittelwerte dabei unüberwacht bestimmt wird. Anhand von zwei Evaluationen, bei denen Ganzkörpersequenzen miteinander verglichen wurden, konnte gezeigt werden, dass diese Strategie bessere Wiedererkennungsergebnisse als konventionelle Multi-Shot-Ansätze erzielt.

Die Ergebnisse der Personenwiedererkennung mittels Tracklet-Deskriptoren konnten durch eine Neusortierung auf Basis des Manifold Learning Algorithmus Laplacian Eigenmaps weiter verbessert werden. Dabei wird eine niedrigdimensionale Repräsentation für die besten n Galerie-Tracklets, die aus dem vorherigen Schritt resultieren, und dem Anfrage-Tracklet gelernt, wobei alle Bildausschnitte, repräsentiert durch Kovarianzdeskriptoren, berücksichtigt werden. Bei diesem Ansatz liegt die Annahme zugrunde, dass die Kovarianzdeskriptoren von Tracklets einer Person jeweils eine Untermannigfaltigkeit im Sym_n^+ bilden.

Darüber hinaus wurden im Rahmen der erscheinungsbasierten Personenwiedererkennung ein mittelwertbasierter Ansatz, der aus mehreren Einzelbildern unterschiedlicher Ansichten einer Person mittelwertbasierte Kovarianzdeskriptoren lernt und eine Fusionsstrategie, die Kovarianzdeskriptoren

mit gelernten Merkmalen verknüpft, auf mehreren Datensätzen evaluiert, um zu untersuchen, inwiefern durch lernbasierte Strategien die handentworfene Ansätze auf Basis der Kovarianzdeskriptoren verbessert werden können. Beide Ansätze erzielten eine Verbesserung im Wiedererkennungsergebnis. Die mittelwertbasierte Strategie erzielte mit *gelernten* anstatt berechneten Mittelwerten leicht bessere Ergebnisse und der Fusionsansatz erzielte durch die Verknüpfung von handentworfenen Kovarianzdeskriptoren mit gelernten Merkmalen bessere Ergebnisse als die jeweils einzelnen eigenständigen Ansätze alleine.

Insgesamt wurde im Rahmen dieser Dissertation ein einheitliches Rahmenwerk erarbeitet, das vier nichtlineare Repräsentationsansätze für Personen in niedrig aufgelösten Videodaten umfasst. Alle Ansätze fokussieren sich auf Videoüberwachungsbilder niedriger Qualität, wobei das Hauptaugenmerk auf der niedrigen Auflösung liegt. Es besteht aus zwei bildsequenzbasierten Personenrepräsentationen, wobei ein Ansatz für die Repräsentation einzelner Personen im Rahmen der Trackinganwendung und der andere sowohl für die kameraübergreifende Rekonstruktion von Personentrajektorien als auch die Suche nach Personen in Videodatensätzen erarbeitet wurde. Im Vergleich zum Stand-der-Forschung werden die Ausreißer dabei besser berücksichtigt. Zudem umfasst es zwei einzelbildbasierte Ansätze, die auf Kovarianzdeskriptormannigfaltigkeiten beruhen, wobei einer davon eine lernbasierte Strategie verfolgt.

9.2 Ausblick

Die unterschiedlichen kovarianzdeskriptorbasierten Strategien, um Personen in niedrig aufgelösten Bildern zu repräsentieren, zeigen vielversprechende Ergebnisse in verschiedenen Anwendungsbereichen. Durch einen einfachen Austausch der Merkmale, die den Kovarianzdeskriptoren zugrunde liegen, können die Ansätze weiter verbessert werden ohne das Bildauswerte-Rahmenwerk anpassen zu müssen. Im Rahmen dieser Arbeit wurden weitere Punkte identifiziert, um das Bildauswerte-Rahmenwerk hinsichtlich bestimmten Aufgabenstellungen und Herausforderungen auszubauen.

Bei hoch aufgelösten Bildern ist die Repräsentation einer Person durch einen einzelnen Kovarianzdeskriptor in der Regel ungeeignet im Vergleich zu vielen anderen handentworfenen Merkmalen und Deskriptoren, da z.B. hoch aufgelöste Texturen, die die Wiedererkennung in der Regel erheblich vereinfachen, nicht ausreichend detailliert repräsentiert werden. Dafür eignen sich, wie beispielsweise in [Bak10] vorgeschlagen, Repräsentationen, die sich aus mehreren Kovarianzdeskriptoren bestimmen, die innerhalb eines Bildausschnitts berechnet wurden. Solche Multi-Kovarianzdeskriptor-Ansätze wären eine nützliche Ergänzung des Rahmenwerks, die zudem auch für die Evaluation auf relevanten öffentlichen Datensätzen interessant sind, welche oft eine höhere als die in dieser Arbeit betrachtete Auflösung haben.

Die vorgestellten Personenrepräsentationen basieren auf Kovarianzdeskriptoren, die schon einige Invarianzen, z.B. gegenüber linearen Beleuchtungsänderungen, bieten. Die Erhöhung von Invarianzen durch Vorverarbeitungsverfahren, wie z.B. Posenschätzungsmethoden, um die Rotationsinvarianz zu erhöhen, wären weitere vielversprechende Ergänzungen des Rahmenwerks. Weitere Verbesserungen für das erscheinungsbasierte Wiedererkennungs- bzw. Suchsystem können durch Verfahren erzielt werden, die anhand von Kontextwissen Galeriedaten vorfiltern oder während der Neusortierung Nutzer-Feedback berücksichtigen. Durch Nutzer-Feedback könnte der Tracklet-Deskriptor durch weitere Mittelwertdeskriptoren ergänzt und damit robuster gemacht werden. Ausführlichere Untersuchungen, inwieweit eine Verbesserung dadurch gegenüber einem einzelnen Mittelwertdeskriptor bzw. wenigen -deskriptoren erzielt werden kann, wären in diesem Kontext von Interesse.

Darüber hinaus sind insbesondere für Suchsysteme lernbasierte Ansätze, wie z.B. die Verfahren in Kapitel 8 vielversprechend. Es wurde gezeigt, dass die Berücksichtigung von Kovarianzdeskriptoren im Trainingsprozess von tiefen künstlichen faltenden neuronalen Netzen die Wiedererkennungsleistung gegenüber einem eigenständigen lernbasierten und handentworfenen Ansatz verbessern kann. Wird ein überwacht gelerntes Suchsystem angestrebt, können solche Fusionsansätze aktuelle Systeme verbessern. Ein *Gesamtnetz*, das sowohl die tiefe Fusion als auch den mittelwertbasierten Ansatz zum Lernen der Kovarianzdeskriptoren in Kapitel 8 integriert, würde eine einheitliche Lösung bieten.

Der mittelwertbasierte Ansatz in Kapitel 8 hat konzeptionell gezeigt, dass eine diskriminativere kovarianzdeskriptorbasierte Repräsentation durch tiefe Netze erzielt werden kann, was vermutlich daran liegt, dass die gelernten Deskriptoren näher an den Mittelwerten liegen, welche die Personen *besser* repräsentieren. Ansätze wie z.B. [Hua16] passen zudem die *euklidische* Architektur der Netze an die riemannsche Mannigfaltigkeit der positiv definiten Kovarianzdeskriptoren an, die es ermöglichen können, diskriminativere Kovarianzdeskriptoren zu lernen. Die Kombination eines solchen Ansatzes mit der tiefen Fusion gelernter Merkmale kann weitere Verbesserungen erzielen. Zudem wären Untersuchungen interessant — ähnlich wie in [Fra17] — inwiefern die Repräsentation gelernter Merkmale durch Kovarianzdeskriptoren — allerdings als Teil des Gesamtnetzes — die Ergebnisse weiter verbessern können.

Literaturverzeichnis

[Ars06] Arsigny, Vincent; Fillard, Pierre; Pennec, Xavier und Ayache, Nicholas: Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine* (2006), Bd. 56(2):S. 411–421

[Ars07] Arsigny, Vincent; Fillard, Pierre; Pennec, Xavier und Ayache, Nicholas: Geometric Means in a Novel Vector Space Structure on Symmetric Positive–Definite Matrices. *SIAM Journal on Matrix Analysis and Applications* (2007), Bd. 29(1):S. 328–347

[Aye12] Ayedi, Walid; Snoussi, Hichem und Abid, Mohamed: A fast multi-scale covariance descriptor for object re-identification. *Pattern Recognition Letters* (2012), Bd. 33(14):S. 1902–1907

[Bak10] Bak, Sławomir; Corvee, Etienne; Brémond, François und Thonnat, Monique: Person Re-identification Using Spatial Covariance Regions of Human Body Parts, in: *Proceedings of the 2010 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, S. 435–440

[Bak11] Bak, Slawomir; Corvee, Etienne; Bremond, Francois und Thonnat, Monique: Multiple-shot human re-identification by Mean Riemannian Covariance Grid, in: *Proceedings of the 2011 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, S. 179–184

[Bak12a] Bak, Sławomir; Charpiat, Guillaume; Corvée, Etienne; Brémond, François und Thonnat, Monique: Learning to Match Appearances by Correlations in a Covariance Metric Space, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Computer Vision – ECCV 2012*, Bd. 7574 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2012), S. 806–820

[Bak12b] Bak, Sławomir; Corvée, Etienne; Brémond, Francois und Thonnat, Monique: Boosted human re-identification using Riemannian manifolds. *Image and Vision Computing* (2012), Bd. 30(6-7):S. 443–452

[Bak14] Bak, Sławomir und Brémond, François: Re-identification by Covariance Descriptors, in: Shaogang Gong; Marco Cristani; Shuicheng Yan und Chen Change Loy (Herausgeber) *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition (ACVPR), Springer London (2014), S. 71–91

[Bat05] Batchelor, P. G.; Moakher, M.; Atkinson, D.; Calamante, F. und Connelly, A.: A rigorous framework for diffusion tensor calculus. *Magnetic Resonance in Medicine* (2005), Bd. 53(1):S. 221–225

[Bau14] Bauml, Martin; Tapaswi, Makarand und Stiefelhagen, Rainer: A time pooled track kernel for person identification, in: *Proceedings of the 2014 International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, S. 7–12

[Baz10] Bazzani, Loris; Cristani, Marco; Perina, Alessandro; Farenzena, Michela und Murino, Vittorio: Multiple-Shot Person Re-identification by HPE Signature, in: *Proceedings of the 2010 International Conference on Pattern Recognition (ICPR)*, S. 1413–1416

[Baz13] Bazzani, Loris; Cristani, Marco und Murino, Vittorio: Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding* (2013), Bd. 117(2):S. 130–144

[Bel03] Belkin, Mikhail und Niyogi, Partha: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* (2003), Bd. 15(6):S. 1373–1396

[Bel04] Belkin, Mikhail und Niyogi, Partha: Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning* (2004), Bd. 56(1-3):S. 209–239

[Ben04] Bengio, Yoshua; Paiement, Jean-François; Vincent, Pascal; Delalleau, Olivier; Le Roux, Nicolas und Ouimet, Marie: Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *Advances in Neural Information Processing Systems* (2004), (16):S. 177–184

[Ben12] Benenson, R.; Mathias, M.; Timofte, R. und van Gool, L.: Pedestrian detection at 100 frames per second, in: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2903–2910

[Ber18] Berger, Marcel: *A Panoramic View of Riemannian Geometry*, Springer-Verlag Berlin Heidelberg, Karlsruhe and Hannover (2018)

[Bew16] Bewley, Alex; Ge, Zongyuan; Ott, Lionel; Ramos, Fabio und Upcroft, Ben: Simple online and realtime tracking, in: *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, S. 3464–3468

[Bey96] Beymer, David und Poggio, Tomaso: Image Representations for Visual Learning. *Science* (1996), Bd. 272(5270):S. 1905–1909

[Bha43] Bhattacharyya, Anil K.: On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* (1943), Bd. 35:S. 99–109

[Bil09] Bilinski, P.; Bremond, F. und Kaaniche, M. B.: Multiple object tracking with occlusions using HOG descriptors and multi resolution images, in: *Proceedings of the 2009 International Conference on Imaging for Crime Detection and Prevention (ICDP)*, S. 1–6

[Bis09] Bishop, Christopher M.: *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer New York, 1. Aufl. (2009)

[Bla03] Black, James; Ellis, Tim und Rosin, Paul: A Novel Method for Video Tracking Performance Evaluation, in: *Proceedings of the 2003 Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, S. 125–132

[Bol10] Bolme, Dav; Beveridge, J. Ross; Draper, Bruce A. und Lui, Yui Man: Visual object tracking using adaptive correlation filters, in: *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2544–2550

[Bou99] Bouguet, Jean-Yves: *Pyramidal implementation of the Lucas Kanade feature tracker*, Technical Report, Intel Corporation, Microprocessor Research Labs (1999)

[Bou09] Bourdev, Lubomir und Malik, Jitendra: Poselets: Body part detectors trained using 3D human pose annotations, in: *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV)*, S. 1365–1372

[Bra12] Brauer, Jürgen; Hübner, Wolfgang und Arens, Michael: Generative 2D and 3D human pose estimation with vote distributions, in: *Proceedings of the 2012 International Symposium on Visual Computing*, S. 470–481

[Bra14] Brauer, Jürgen: *Human pose estimation with implicit shape models: Zugl.: Karlsruhe, KIT, Diss., 2014*, Bd. 6 von *Schriftenreihe Automatische Sichtprüfung und Bildverarbeitung*, KIT Scientific Publishing, Karlsruhe (2014)

[Bro93] Bromley, J. et al.: Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence* (1993), Bd. 07(04):S. 669–688

[Cam08] Camastra, Francesco und Vinciarelli, Alessandro: *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*, Advanced Information and Knowledge Processing, Springer London, 1. Aufl. (2008)

[CEN96] CENELEC: Alarm systems - CCTV surveillance systems for use in security applications, part 7, EN50132-7 (1996)

[Che11] Cheng, Dong Seon; Cristani, Marco; Stoppa, Michele; Bazzani, Loris und Murino, Vittorio: Custom Pictorial Structures for Re-identification, in: Jesse Hoey; Stephen McKenna; Emanuele Trucco und Jianguo Zhang (Herausgeber) *British Machine Vision Conference 2011*, Bd. 25.68, S. 1–11

[Che15] Chen, Ying-Cong; Zheng, Wei-Shi und Lai, Jianhuang: Mirror representation for modeling view-specific transform in person re-identification, in: Qiang Yang und Michael J. Wooldridge (Herausgeber) *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, AAAI Press International Joint Conferences on Artificial Intelligence, Palo Alto, California (2015), S. 3402–3408

[Che17] Chen, Weihua; Chen, Xiaotang; Zhang, Jianguo und Huang, Kaiqi: Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification, in: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1320–1329

[Cho13] Choudhury, Sruti D. und Tjahjadi, Tardi: Gait recognition based on shape and motion analysis of silhouette contours. *Computer Vision and Image Understanding* (2013), Bd. 117(12):S. 1770–1785

[Cho17] Choi, Jongwon; Chang, Hyung Jin; Yun, Sangdoo; Fischer, Tobias; Demiris, Yiannis und Choi, Jin Young: Attentional Correlation Filter Network for Adaptive Visual Tracking, in: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 4828–4837

[Chu95] Chung, Fan R. K.: *Spectral Graph Theory*, Bd. 92 von *CBMS Lecture Notes*, American Mathematical Society (1995)

[Com02] Comaniciu, D. und Meer, P.: Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2002), Bd. 24(5):S. 603–619

[Con09] Cong, D.N.T.; Achard, C.; Khoudour, L. und Douadi, L.: Video Sequences Association for People Re-identification Across Multiple Non-overlapping Cameras, in: *Proceedings of the 2009 International Conference on Image Analysis and Processing (ICIAP)*, S. 179–189

[Con13] Conde, Cristina; Moctezuma, Daniela; Martín De Diego, Isaac und Cabello, Enrique: HoGG: Gabor and HoG-based human detection for surveillance in non-controlled environments. *Neurocomputing* (2013), Bd. 100:S. 19–30

[Cos14] Costea, Arthur Daniel und Nedevschi, Sergiu: Word Channel Based Multiscale Pedestrian Detection without Image Resizing and Using Only One Classifier, in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2393–2400

[Cro84] Crow, Franklin C.: Summed-area tables for texture mapping, in: *Proceedings of the 1984 Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, S. 207–212

[Csu04] Csurka, G.; Dance, C. R.; Fan, L.; Willamowski, J. und Bray, C.: Visual categorization with bags of keypoints, in: Tomas Pajdla und Jiri Matas (Herausgeber) *Proceedings of the 2004 European Conference on Computer Vision (ECCV) - Workshops*, Lecture Notes in Computer Science (LNCS), Springer Berlin Heidelberg (2004), S. 1–22

[D'A11] D'Angelo, Angela und Dugelay, Jean-Luc: People re-identification in camera networks based on probabilistic color histograms, in: Amir Said; Onur G. Guleryuz und Robert L. Stevenson (Herausgeber) *Proceedings of the 2011 SPIE Electronic Imaging; Visual Information Processing and Communication II*, Bd. 7882 von *SPIE Proceedings*, S. 23–27

[Dal05] Dalal, N. und Triggs, B.: Histograms of Oriented Gradients for Human Detection, in: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 886–893

[Dal06] Dalal, N.; Triggs, B. und Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance, in: Aleš Leonardis; Horst Bischof und Axel Pinz (Herausgeber) *Computer vision– ECCV 2006*, Bd. 3951-3954 von *Lecture Notes in Computer Science*, Springer, Berlin and New York (2006), S. 428–441

[Dan15] Danelljan, Martin; Hager, Gustav; Khan, Fahad Shahbaz und Felsberg, Michael: Convolutional Features for Correlation Filter Based Visual Tracking, in: *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, S. 621–629

[Dan16] Danelljan, M.; Robinson, A.; Shahbaz Khan, F. und Felsberg, M.: Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking, in: Bastian Leibe; Jiri Matas; Nicu Sebe und Max Welling (Herausgeber) *Computer vision - ECCV 2016*, Bd. 9909 von *Lecture Notes in Computer Science*, Springer, Cham (2016)

[Dao17] Dao Vu, Quang und Chung, Sun-Tae: Real-time robust human tracking based on Lucas-Kanade optical flow and deep detection for embedded surveillance, in: *Proceedings of the 2017 International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, S. 1–6

[Den09] Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li., K. und Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, in: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, S. 248–255

[Dol09] Dollar, Piotr; Tu, Zhuowen; Perona, Pietro und Belongie, Serge: Integral Channel Features, in: A. Cavallaro; S. Prince und D. Alexander (Herausgeber) *Proceedings of the 2009 British Machine Vision Conference (BMVC)*, S. 91.1

[dR03] de Ridder, Dick; Kouropteva, Olga; Okun, Oleg; Pietikäinen, Matti und Duin, Robert P. W.: Supervised Locally Linear Embedding, in: Okyay Kaynak; Ethem Alpaydin; Erkki Oja und Lei Xu (Herausgeber) *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP*, Bd. 2714 von *Lecture*

Notes in Computer Science (LNCS), Springer Berlin Heidelberg (2003), S. 333–341

[dW04] DE WINTER, Joeri und WAGEMANS, Johan: Contour-based object identification and segmentation: Stimuli, norms and data, and software tools. *Behavior Research Methods, Instruments and Computers* (2004), Bd. 36(4):S. 604–624

[Eis14] EISELEIN, Volker; STERNHARZ, Gleb; SENST, Tobias; KELLER, Ivo und SIKORA, Thomas: Person Re-identification Using Region Covariance in a Multi-feature Approach, in: Aurélio Campilho und Mohamed Kamel (Herausgeber) *Image Analysis and Recognition*, Bd. 8815 von *Lecture Notes in Computer Science*, Springer International Publishing (2014), S. 77–84

[Elg04] ELGAMMAL, Ahmed und LEE, Chan-Su: Inferring 3D body pose from silhouettes using activity manifold learning, in: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 681–688

[Elg08] ELGAMMAL, Ahmed und LEE, Chan-Su: The Role of Manifold Learning in Human Motion Analysis, in: Bodo Rosenhahn; Reinhard Klette und Dimitris Metaxas (Herausgeber) *Human Motion*, Bd. 36 von *Computational Imaging and Vision*, Springer Netherlands (2008), S. 25–56

[Far10] FARENZENA, M.; BAZZANI, L.; PERINA, A.; MURINO, V. und CRISTANI, M.: Person re-identification by symmetry-driven accumulation of local features, in: *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2360–2367

[Fau15] FAULKNER, Hayden; SHEHU, Ergnoor; SZPAK, Zygmunt L.; CHOJNACKI, Wojciech; TAPAMO, Jules R.; DICK, Anthony und VAN DEN HENGEL, Anton: A Study of the Region Covariance Descriptor: Impact of Feature Selection and Image Transformations, in: *Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, S. 1–8

[Fle04] FLETCHER, P. Thomas und JOSHI, Sarang: Principal Geodesic Analysis on Symmetric Spaces: Statistics of Diffusion Tensors, in:

Milan Sonka; Ioannis A. Kakadiaris und Jan Kybic (Herausgeber) *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*, Bd. 3117 von *Lecture Notes in Computer Science (LNCS)*, Springer Berlin Heidelberg (2004), S. 87–98

[Fle07] Fletcher, P. Thomas und Joshi, Sarang: Riemannian geometry for the statistical analysis of diffusion tensor data. *International Journal of Signal Processing* (2007), Bd. 87(2):S. 250–262

[För99] Förstner, Wolfgang und Moonen, Boudewijn: *A Metric for Covariance Matrices*, Technical Report 1999.6, Department of Geodesy and Geoinformatics, Stuttgart University (1999)

[For07] Forssen, Per-Erik: Maximally Stable Colour Regions for Recognition and Matching, in: *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8

[Fra17] Franco, Alexandre und Oliveira, Luciano: Convolutional covariance features: Conception, integration and performance in person re-identification. *Pattern Recognition* (2017), Bd. 61:S. 593–609

[Fuk80] Fukushima, Kunihiko: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* (1980), Bd. 36(4):S. 193–202

[Gav07] Gavrila, Dariu M. und Munder, Stefan: Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. *International Journal of Computer Vision* (2007), Bd. 73(1):S. 41–59

[Gha15] Ghasemi, Abouzar und Kumar, C.N Ravi: A Survey of Multi Object Tracking and Detecting Algorithm in Real Scene use in video surveillance systems. *International Journal of Computer Trends and Technology* (2015), Bd. 29(1):S. 31–39

[Ghe06] Gheissari, N.; Sebastian, T. B. und Hartley, R.: Person Reidentification Using Spatiotemporal Appearance, in: *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 2, S. 1528–1535

[Gir14] Girshick, Ross; Donahue, Jeff; Darrell, Trevor und Malik, Jitendra: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 580–587

[Glo17] Global Aerospace & Defense Research Team at Frost & Sullivan: United States Video Surveillance Market, Forecast to 2023: Smarter Security and Surveillance Technology Enhances Market Growth (Oktober 2017)

[Gon14] Gong, Yunchao; Wang, Liwei; Guo, Ruiqi und Lazebnik, Svetlana: Multi-scale Orderless Pooling of Deep Convolutional Activation Features, in: David Fleet; Tomas Pajdla; Bernt Schiele und Tinne Tuytelaars (Herausgeber) *Computer Vision – ECCV 2014*, Bd. 8695 von *Lecture Notes in Computer Science*, Springer International Publishing (2014), S. 392–407

[Goo16] Goodfellow, Ian; Bengio, Yoshua und Courville, Aaron: *Deep learning*, MIT Press, Cambridge, Massachusetts and London, England (2016), URL `http://www.deeplearningbook.org/`

[Gor17] Gordon, D.; Farhadi, A. und Fox, D.: Re3: Real-Time Recurrent Regression Networks for Visual Tracking of Generic Objects. *arXiv:1705.06368* (2017)

[Gra07] Gray, D.; Brennan, S. und Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking, in: *Proceedings of the 2007 IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)* (2007)

[Gra08] Gray, D. und Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features, in: David Forsyth; Philip Torr und Andrew Zisserman (Herausgeber) *Computer vision - ECCV 2008*, Bd. 5302 von *Lecture Notes in Computer Science*, Springer, Berlin (2008)

[Gri09] Grigoryan, Alexander: *Heat Kernel and Analysis on Manifolds*, Bd. 47 von *AMS/IP Studies in Advanced Mathematics*, American Mathematical Society (2009)

[Gri18] Grinberg, Michael: *Feature-Based Probabilistic Data Association for Video-Based Multi-Object Tracking: (noch nicht veröffentlicht): Zugl.: Karlsruhe, KIT, Diss., 2018*, Karlsruher Schriften zur Anthropomatik, KIT Scientific Publishing and Technische Informationsbibliothek u. Universitätsbibliothek, Karlsruhe and Hannover (2018)

[Hah04] Hahnel, M.; Klunder, D. und Kraiss, K.-F.: Color and texture features for person recognition, in: *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IJCNN)*, S. 647–652

[Har79] Hartigan, J. A. und Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* (1979), Bd. 28(1):S. 100–108

[Har11] Hare, Sam; Saffari, Amir und Torr, Philip H. S.: Struck: Structured output tracking with kernels, in: *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, S. 263–270

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing und Sun, Jian: Deep Residual Learning for Image Recognition, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 770–778

[Hen11] Henriques, Joao F.; Caseiro, Rui und Batista, Jorge: Globally optimal solution to multi-object tracking with merged measurements, in: *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, S. 2470–2477

[Hen15] Henriques, João F.; Caseiro, Rui; Martins, Pedro und Batista, Jorge: High-Speed Tracking with Kernelized Correlation Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2015), Bd. 37(3):S. 583–596

[Her11] Herrmann, Christian; Manger, Daniel und Metzler, Juergen: Feature-based localization refinement of players in soccer using plausibility maps, in: *Proceedings of the 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*

[Her16] Herrmann, Christian; Müller, Thomas; Willersinn, Dieter und Beyerer, Jürgen: Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs, in: David A. Huckridge; Reinhard Ebert und Stephen T. Lee (Herausgeber) *Proceedings of the 2016 SPIE Security + Defence*, SPIE Proceedings, SPIE, S. 99870I

[Her17] Hermans, A.; Beyer, L. und Leibe, B.: In Defense of the Triplet Loss for Person Re-Identification. *arXiv:1703.07737* (2017)

[Her18] Herrmann, Christian: *Video-to-Video FaceRecognition for Low-QualityVideo-to-Video Face Recognition for Low-Quality Surveillance Data: (noch nicht veröffentlicht): Zugl.: Karlsruhe, KIT, Diss., 2018*, Karlsruher Schriften zur Anthropomatik, KIT Scientific Publishing and Technische Informationsbibliothek u. Universitätsbibliothek, Karlsruhe and Hannover (2018)

[Hir11] Hirzer, M.; Beleznai, C.; Roth, P. M. und Bischof, H.: Person Re-identification by Descriptive and Discriminative Classification, in: Anders Heyden und Fredrik Kahl (Herausgeber) *Image analysis*, Bd. 6688 von *Lecture Notes in Computer Science*, Springer, Berlin (2011), S. 91–102

[Hon10] Hong, Xiaopeng; Chang, Hong; Shan, Shiguang; Zhong, Bineng; Chen, Xilin und Gao, Wen: Sigma Set Based Implicit Online Learning for Object Tracking. *IEEE Signal Processing Letters* (2010), Bd. 17(9):S. 807–810

[Hon15] Hong, S.; You, T.; Kwak, S. und Han, B.: Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network, in: *Proceedings of the 2015 International Conference on Machine Learning (ICML)*, Bd. 37, S. 597–606

[Hor85] Horn, Roger A. und Johnson, Charles R.: *Matrix analysis*, Cambridge University Press, 1 Aufl. (1985)

[Hor17] Horev, Inbal; Yger, Florian und Sugiyama, Masashi: Geometry-aware principal component analysis for symmetric positive definite matrices. *Machine Learning* (2017), Bd. 106(4):S. 493–522

[Hu12] Hu, Weiming; Li, Xi; Luo, Wenhan; Zhang, Xiaoqin; Maybank, Stephen und Zhang, Zhongfei: Single and multiple object tracking using log-euclidean Riemannian subspace and block-division appearance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2012), Bd. 34(12):S. 2420–2440

[Hua16] Huang, Zhiwu und van Gool, Luc: A Riemannian Network for SPD Matrix Learning. *arXiv:1608.04233* (2016)

[Hüb08] Hübner, Yvonne; Metzler, Jürgen; Dürr, Bernhard; Jäger, Uwe und Willersinn, Dieter: Assessment and optimization of methods for tracking people in riot control scenarios, in: Gary W. Kamerman; Ove K. Steinvall und et al. (Herausgeber) *Proceedings of the 2008 SPIE Europe Security and Defence*, Bd. 7114 von *SPIE Proceedings*, S. 711403

[Hus10] Hussain, Sibt ul und Triggs, Bill: Feature Sets and Dimensionality Reduction for Visual Object Detection, in: Frédéric Labrosse; Reyer Zwiggelaar; Yonghuai Liu und Bernie Tiddeman (Herausgeber) *Proceedings of the 2010 British Machine Vision Conference (BMVC)*, S. 112.1–112.10

[Ily10] Ilyas, Atif; Scuturici, Mihaela und Miguet, Serge: Intercamera color calibration for object re-identification and tracking, in: *Proceedings of the 2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, S. 188–193

[Jäg08] Jäger, Uwe; Höpken, Marc; Dürr, Bernhard; Metzler, Jürgen und Willersinn, Dieter: Multisensor benchmark data for riot control, in: Gary W. Kamerman; Ove K. Steinvall und et al. (Herausgeber) *Proceedings of the 2008 SPIE Europe Security and Defence*, SPIE Proceedings, S. 711403

[Jep03] Jepson, A. D.; Fleet, D. J. und El-Maraghi, T. F.: Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2003), Bd. 25(10):S. 1296–1311

[Jia09] Jiang, Quansheng und Jia, Minping: Supervised Laplacian Eigenmaps for Machinery Fault Classification, in: *Proceedings of the*

2009 WRI World Congress on Computer Science and Information Engineering, S. 116–120

[Jol86] Jolliffe, I. T.: *Principal Component Analysis*, Springer Series in Statistics, Springer, New York, NY (1986)

[Jul96] Julier, S. und Uhlmann, J. K.: *A general method for approximating non-linear transformations of probability distributions*, Technical Report, University of Oxford, United Kingdom, Oxford (1996)

[Jün11a] Jüngling, Kai: *Ein generisches System zur automatischen Detektion, Verfolgung und Wiedererkennung von Personen in Videodaten*, Dissertation, KIT, Karlsruhe (2011)

[Jün11b] Jüngling, Kai und Arens, Michael: Local Feature Based Person Detection and Tracking Beyond the Visible Spectrum, in: Riad Hammoud; Guoliang Fan; Robert W. McMillan und Katsushi Ikeuchi (Herausgeber) *Machine Vision Beyond Visible Spectrum*, Bd. 1 von *Augmented Vision and Reality*, Springer Berlin Heidelberg (2011), S. 3–32

[Kar77] Karcher, Hermann: Riemannian centre of mass and mollifier smoothing. *Communications in Pure and Applied Mathematics* (1977), Bd. 30:S. 509–541

[Kar15] Karanam, Srikrishna; Li, Yang und Radke, Richard J.: Sparse re-id: Block sparsity for person re-identification, in: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, S. 33–40

[Kau15] Kaur, Ramandeep und Himanshi, Er.: Face recognition using Principal Component Analysis, in: *Proceedings of the 2015 IEEE International Advance Computing Conference (IACC)*, S. 585–589

[Kaw12] Kawai, Ryo; Makihara, Yasushi; Hua, Chunsheng; Iwama, Haruyuki und Yagi, Yasushi: Person re-identification using view-dependent score-level fusion of gait and color features, in: *Proceedings of the 2012 IEEE International Conference on Pattern Recognition (ICPR)*, S. 2694–2697

[Ken90] Kendall, Wilfrid S.: Probability, Convexity, and Harmonic Maps with Small Image I: Uniqueness and Fine Existence. *Proceedings of the London Mathematical Society* (1990), Bd. 61(2):S. 371–406

[Klu10] Kluckner, Stefan; Mauthner, Thomas; Roth, Peter M. und Bischof, Horst: Semantic Classification in Aerial Imagery by Integrating Appearance and Height Information, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Computer Vision – ACCV 2009*, Bd. 5995 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2010), S. 477–488

[Kos12] Kostinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M. und Bischof, H.: Large scale metric learning from equivalence constraints, in: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2288–2295

[Kou03] Kouropteva, Olga; Okun, Oleg und Pietikäinen, Matti: Supervised Locally Linear Embedding Algorithm for Pattern Recognition, in: Francisco José Perales; Aurélio J. C. Campilho und et al. La Blanca (Herausgeber) *Pattern Recognition and Image Analysis*, Bd. 2652 von *Lecture Notes in Computer Science (LNCS)*, Springer Berlin Heidelberg (2003), S. 386–394

[Kri12] Krizhevsky, Alex; Sutskever, Ilya und Hinton, Geoffrey E.: ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems* (2012), Bd. 25:S. 1097–1105

[Kri15] Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Cehovin, L.; Fernandez, G. und et al.: The Visual Object Tracking VOT2015 Challenge Results, in: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) - Workshop*, S. 564–586

[Kri16] Kristan M. et al.: The Visual Object Tracking VOT2016 Challenge Results, in: Gang Hua und Hervé Jégou (Herausgeber) *Computer Vision - ECCV 2016 Workshops*, Bd. 9914 von *Lecture Notes in Computer Science*, Springer International Publishing, Cham and s.l. (2016)

[Kuh55] Kuhn, H. W.: The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* (1955), Bd. 2(1-2):S. 83–97

[Kvi13] Kviatkovsky, Igor; Adam, Amit und Rivlin, Ehud: Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013), Bd. 35(7):S. 1622–1634

[Lay12] Layne, Ryan; Hospedales, Tim und Gong, Shaogang: Person Re-identification by Attributes, in: *Proceedings of the 2012 British Machine Vision Conference (BMVC)*, S. 1–11

[Lee04] Lee, D. .J.; Zhan, P.; Thomas, A. und Schoenberger, R.: Shape-based Human Intrusion Detection, in: Zia-ur Rahman; Robert A. Schowengerdt und Stephen E. Reichenbach (Herausgeber) *Proceedings of the 2004 International Symposium on Defense, Security and Sensing; Visual Information Processing XIII*, Bd. 5438 von *SPIE Proceedings*, S. 81–91

[Lee07] Lee, John A. und Verleysen, Michel: *Nonlinear Dimensionality Reduction*, Information Science and Statistics, Springer New York, 1 Aufl. (2007)

[Lei04] Leibe, Bastian; Leonardis, Ales und Schiele, Bernt: Combined Object Categorization and Segmentation With An Implicit Shape Model, in: Tomas Pajdla und Jiri Matas (Herausgeber) *Proceedings of the 2004 European Conference on Computer Vision (ECCV)*, Bd. 3024 von *Lecture Notes in Computer Science (LNCS)*, Springer Berlin Heidelberg (2004), S. 17–32

[Lei06] Leibe, Bastian; Leonardis, Ales und Schiele, Bernt: An Implicit Shape Model for Combined Object Categorization and Segmentation, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Toward Category-Level Object Recognition*, Bd. 4170 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2006), S. 508–524

[Len04] Lenglet, Christophe; Rousson, Mikaël; Deriche, Rachid und Faugeras, Olivier: *StatisticsToward Segmentation of 3D Probability Density Fields by Surface Evolution: Application to Diffusion MRI*, Research Report 5243, INRIA (2004)

[Len06] Lenglet, Christophe; Rousson, Mikaël; Deriche, Rachid und Faugeras, Olivier: Statistics on the Manifold of Multivariate Normal Distributions: Theory and Application to Diffusion Tensor MRI Processing. *Journal of Mathematical Imaging and Vision* (2006), Bd. 25(3):S. 423–444

[Lev02] Levin, Anat und Shashua, Amnon: Principal Component Analysis over Continuous Subspaces and Intersection of Half-Spaces, in: Gerhard Goos; Juris Hartmanis; Jan van Leeuwen und et al. (Herausgeber) *Computer Vision — ECCV 2002*, Bd. 2352 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2002), S. 635–650

[Li08] Li, Xi; Hu, Weiming; Zhang, Zhongfei; Zhang, Xiaoqin; Zhu Mingliang und Cheng, Jian: Visual tracking via incremental Log-Euclidean Riemannian subspace learning, in: *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8

[Li12a] Li, Peihua und Wang, Qilong: Local Log-Euclidean Covariance Matrix (L2ECM) for Image Representation and Its Applications, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Computer Vision – ECCV 2012*, Bd. 7574 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2012), S. 469–482

[Li12b] Li, Wei; Zhao, Rui und Wang, Xiaogang: Human Reidentification with Transferred Metric Learning, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Proceedings of the 2012 Asian Conference on Computer Vision (ACCV)*, Bd. 7724 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg (2012), S. 31–44

[Li14] Li, Xiao; Song, Mingli; Tao, Dacheng; Zhou, Xingchen; Chen, Chun und Bu, Jiajun: Semi-supervised Coupled Dictionary Learning for Person Re-identification, in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 3550–3557

[Li15] Li, Bo; Li, Jun und Zhang, Xiao-Ping: Nonparametric discriminant multi-manifold learning for dimensionality reduction. *Neurocomputing* (2015), Bd. 152:S. 121–126

[Lia05] Liang, Dong; Yang, Jie; Zheng, Zhonglong und Chang, Yuchou: A facial expression recognition system based on supervised locally linear embedding. *Pattern Recognition Letters* (2005), Bd. 26(15):S. 2374–2389

[Lia11] Liao, Wen-Hung und Huang, Ling-Wei: Pedestrian Detection Using Covariance Descriptor and On-line Learning, in: *Proceedings of the 2011 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, S. 179–182

[Lia15] Liao, Shengcai; Hu, Yang; Xiangyu Zhu und Li, Stan Z.: Person re-identification by Local Maximal Occurrence representation and metric learning, in: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2197–2206

[Liu12] Liu, Chunxiao; Gong, Shaogang; Loy, Chen Change und Lin, Xinggang: Person Re-identification: What Features Are Important?, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Bd. 7583 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2012), S. 391–401

[Liu16] Liu, Wei; Anguelov, Dragomir; Erhan, Dumitru; Szegedy, Christian; Reed, Scott; Fu, Cheng-Yang und Berg, Alexander C.: SSD: Single Shot MultiBox Detector, in: Bastian Leibe; Jiri Matas; Nicu Sebe und Max Welling (Herausgeber) *Computer Vision - ECCV 2016*, Bd. 9905 von *Lecture Notes in Computer Science*, Springer, Cham (2016), S. 21–37

[Low04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* (2004), Bd. 60(2):S. 91–110

[Ma12a] Ma, Bingpeng; Su, Yu und Jurie, Frederic: BiCov: a novel image representation for person re-identification and face verification, in:

R. Bowden; J. Collomosse und K. Mikolajczyk (Herausgeber) *Proceedings of the 2012 British Machine Vision Conference (BMVC)*, S. 57.1

[Ma12b] Ma, Bingpeng; Su, Yu und Jurie, Frédéric: Local Descriptors Encoded by Fisher Vectors for Person Re-identification, in: David Hutchison; Takeo Kanade; Josef Kittler und et al. (Herausgeber) *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Bd. 7583 von *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg (2012), S. 413–422

[Ma15] Ma, Chao; Huang, Jia-Bin; Yang, Xiaokang und Yang, Ming-Hsuan: Hierarchical Convolutional Features for Visual Tracking, in: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, S. 3074–3082

[Mad07] Madden, Christopher; Cheng, Eric Dahai und Piccardi, Massimo: Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications* (2007), Bd. 18(3-4):S. 233–247

[Mar15] Marciniak, Tomasz; Chmielewska, Agata; Weychan, Radoslaw; Parzych, Marianna und Dabrowski, Adam: Influence of low resolution of images on reliability of face detection and recognition. *Multimedia Tools and Applications* (2015), Bd. 74(12):S. 4329–4349

[Mat16] Matsukawa, Tetsu; Okabe, Takahiro; Suzuki, Einoshin und Sato, Yoichi: Hierarchical Gaussian Descriptor for Person Re-identification, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1363–1372

[Mat17] Matsuzawa, Tomoki; Ito, Eisuke; Relator, Raissa; Sese, Jun und Kato, Tsuyoshi: Stochastic Dykstra Algorithms for Distance Metric Learning with Covariance Descriptors. *IEICE Transactions on Information and Systems)* (2017), Bd. E100.D(4):S. 849–856

[McL16] McLaughlin, Niall; Rincon, Jesus Martinez del und Miller, Paul: Recurrent Convolutional Network for Video-Based Person Re-identification, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1325–1334

[Met07] Metzler, Jürgen und Willersinn, Dieter: Toward a sensor-based threat warning system for patrols in MOUT scenarios, in: Firooz A. Sadjadi (Herausgeber) *Proceedings of the 2007 International Symposium on Defense, Security and Sensing; Automatic Target Recognition XVII*, Bd. 6566 von *Proceedings of SPIE*, S. 65660T–1 – 65660T–9

[Met09] Metzler, Jürgen und Willersinn, Dieter: Robust tracking of people in crowds with covariance descriptors, in: Zia-ur Rahman; Stephen E. Reichenbach und Mark A. Neifeld (Herausgeber) *Proceedings of the 2009 SPIE Defense, Security, and Sensing*, Bd. 7341 von *SPIE Proceedings*, S. 73410T

[Met10] Metzler, Jürgen und Willersinn, Dieter: Human detection in MOUT scenarios using covariance descriptors and supervised manifold learning, in: Zia-ur Rahman; Stephen E. Reichenbach und Mark A. Neifeld (Herausgeber) *Proceedings of the 2010 International Symposium on Defense, Security and Sensing; Visual Information Processing XIX*, Bd. 7701 von *SPIE Proceedings*, S. 67942K–1 – 67942K–6

[Met12a] Metzler, Jurgen: Appearance-Based Re-identification of Humans in Low-Resolution Videos Using Means of Covariance Descriptors, in: *Proceedings of the 2012 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, S. 191–196

[Met12b] Metzler, Jurgen: Two-stage appearance-based re-identification of humans in low-resolution videos, in: *Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, S. 19–24

[Met14] Metzler, Jürgen; Monari, Eduardo und Kuntzsch, Colin: Application-driven merging and analysis of person trajectories for distributed smart camera networks, in: Robert P. Loce und Eli Saber (Herausgeber) *Proceedings of the 2014 IS&T/SPIE Electronic Imaging*, SPIE Proceedings, S. 90260I

[Mig12] Mignon, A. und Jurie, F.: PCCA: A new approach for distance learning from sparse pairwise constraints, in: *Proceedings of the*

2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), S. 2666–2672

[Moa05] Moakher, Maher: A Differential Geometric Approach to the Geometric Mean of Symmetric Positive-Definite Matrices. *SIAM Journal on Matrix Analysis and Applications* (2005), Bd. 26(3):S. 735–747

[Mon11] Monari, Eduardo: *Dynamische Sensorselektion zur auftragsorientierten Objektverfolgung in Kameranetzwerken: Zugl.: Karlsruhe, KIT, Diss., 2010*, Bd. 8 von *Karlsruher Schriften zur Anthropomatik*, KIT Scientific Publ and Technische Informationsbibliothek u. Universitätsbibliothek, Karlsruhe and Hannover (2011), URL `http://edok01.tib.uni-hannover.de/edoks/e01fn12/680096477.pdf`

[Mon13] Monari, Eduardo: Illumination Invariant Background Subtraction for Pan/Tilt Cameras using DoG Responses, in: *Proceedings of the 2013 International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, S. 122–128

[Mue17] Mueller, Matthias; Smith, Neil und Ghanem, Bernard: Context-Aware Correlation Filter Tracking, in: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1387–1395

[Mül08] Müller, Thomas und Müller, Markus: CACAMO - computer-aided camouflage assessment of moving objects, in: Gerald C. Holst (Herausgeber) *Proceedings of the 2008 International Symposium on Defense, Security and Sensing; Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XIX*, Bd. 69410V von *Proceedings of SPIE*, S. 69410V–1 – 69410V–12

[Mül11] Müller, Thomas; Manger, Daniel und Metzler, Jürgen: Recognition of soccer players after occlusions using temporal color signatures, in: *Proceedings of the 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*

[Nam16] Nam, Hyeonseob und Han, Bohyung: Learning Multi-domain Convolutional Neural Networks for Visual Tracking, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 4293–4302

[Ngu16] Nguyen, Duc Thanh; Li, Wanqing und Ogunbona, Philip O.: Human detection from images and videos: A survey. *Pattern Recognition* (2016), Bd. 51:S. 148–175

[Oja02] Ojala, T.; Pietikainen, M. und Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2002), Bd. 24(7):S. 971–987

[Oli16] Oliveira, Gabriel L.; Valada, Abhinav; Bollen, Claas; Burgard, Wolfram und Brox, Thomas: Deep learning for human part discovery in images, in: *Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, S. 1634–1641

[Pai07] Paisitkriangkrai, Sakrapee; Shen, Chunhua und Zhang, Jian: An Experimental Evaluation of Local Features for Pedestrian Classification, in: *Proceedings of the 2007 Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA)*, S. 53–60

[Pal08] Palaio, Helio und Batista, Jorge: Multi-object tracking using an adaptive transition model particle filter with region covariance data association, in: *Proceedings of the 2008 International Conference on Pattern Recognition (ICPR)*, S. 1–4

[Pea01] Pearson, Karl: On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (1901), Bd. 6(2):S. 559–572

[Ped13] Pedagadi, Sateesh; Orwell, James; Velastin, Sergio und Boghossian, Boghos: Local Fisher Discriminant Analysis for Pedestrian Re-identification, in: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 3318–3325

[Pen99] Pennec, Xavier: Probabilities And Statistics On Riemannian Manifolds: Basic Tools For Geometric Measurements, in: *Proceedings of the 1999 IEEE Workshop on Nonlinear Signal and Image Processing (NSIP)*, Bd. 1, S. 194–198

[Pen06a] Pennec, Xavier: Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision* (2006), Bd. 25(1):S. 127–154

[Pen06b] Pennec, Xavier: *Statistical Computing on Manifolds for Computational Anatomy*, Habilitationsschrift, Université Nice Sophia Antipolis (2006)

[Pen06c] Pennec, Xavier; Fillard, Pierre und Ayache, Nicholas: A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision* (2006), Bd. 66(1):S. 41–66

[Per07] Perronnin, Florent und Dance, Christopher: Fisher Kernels on Visual Vocabularies for Image Categorization, in: *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8

[Pog90] Poggio, T. und Girosi, F.: Networks for approximation and learning. *Proceedings of the IEEE* (1990), Bd. 78(9):S. 1481–1497

[Por05] Porikli, F.: Integral histogram: a fast way to extract histograms in Cartesian spaces, in: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 829–836

[Por06a] Porikli, Fatih und Tuzel, Oncel: Fast Construction of Covariance Matrices for Arbitrary Size Image Windows, in: *Proceedings of the 2006 International Conference on Image Processing (ICIP)*, S. 1581–1584

[Por06b] Porikli, Fatih; Tuzel, Oncel und Meer, Peter: Covariance Tracking using Model Update Based on Means on Riemannian Manifolds, in: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 728–735

[Qu18] Qu, Chengchao; Metzler, Jürgen und Monari, Eduardo;: ivisX: an Integrated Video Investigation Suite for Forensic Applications, in: *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); Cross-Domain Biometric Recognition Workshop*

[Raw17] Rawat, Waseem und Wang, Zenghui: Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation* (2017), Bd. 29(9):S. 2352–2449

[Rez17] Rezaei, Mahdi und Klette, Reinhard: *Computer Vision for Driver Assistance*, Bd. 45, Springer International Publishing, Cham (2017)

[Rot08] Roth, Peter M. und Winter Martin: *Survey of appearance-based methods for object recognition*, Technical Report ICG-TR-01/08, Massachusetts Institute of Technology (2008)

[Row00] Roweis, Sam T. und Saul, Lawrence K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* (2000), Bd. 290(5500):S. 2323–2326

[Sag14] Saghafi, Mohammad Ali; Hussain, Aini; Zaman, Halimah Badioze und Saad, Mohamad Hanif Md.: Review of person re-identification techniques. *IET Computer Vision* (2014), Bd. 8(6):S. 455–474

[Sah16] Sahbani, Bima und Adiprawita, Widyawardana: Kalman filter and Iterative-Hungarian Algorithm implementation for low complexity point tracking as part of fast multiple object tracking system, in: *Proceedings of the 2016 International Conference on System Engineering and Technology (ICSET)*, S. 109–115

[Sal12] Salti, Samuele; Cavallaro, Andrea und Di Stefano, Luigi: Adaptive Appearance Modeling for Video Tracking: Survey and Evaluation. *IEEE Transactions on Image Processing* (2012), Bd. 21(10):S. 4334–4348

[San13] Sanin, Andres; Sanderson, Conrad; Harandi, Mehrtash T. und Lovell, Brian C.: Spatio-temporal covariance descriptors for action and gesture recognition, in: *Proceedings of the 2013*

IEEE Workshop on Applications of Computer Vision (WACV), S. 103–110

[Sat12] Satta, Riccardo; Fumera, Giorgio und Roli, Fabio: Appearance-based people recognition by local dissimilarity representations, in: *Proceedings of the 2012 ACM Workshop on Multimedia and Security*, S. 151–156

[Sat14] Satpathy, Amit; Jiang, Xudong und Eng, How-Lung: Human detection by quadratic classification on subspace of extended histogram of gradients. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* (2014), Bd. 23(1):S. 287–297

[Sch02] Schölkopf, Bernhard und Smola, Alexander J.: *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, Adaptive computation and machine learning, MIT Press, Cambridge, Mass (2002), URL `http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=78092`

[Sch15] Schumann, Arne und Stiefelhagen, Rainer: Transferring attributes for person re-identification, in: *Proceedings of the 2015 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, S. 1–6

[Sch17] Schumann, Arne und Metzler, Jürgen: Person re-identification across aerial and ground-based cameras by deep feature fusion, in: Firooz A. Sadjadi und Abhijit Mahalanobis (Herausgeber) *Proceedings of the 2017 SPIE Defense + Security*, SPIE Proceedings, S. 102020A

[Ser14] Serra, Giuseppe; Grana, Costantino; Manfredi, Marco und Cucchiara, Rita: Covariance of Covariance Features for Image Classification, in: Joemon Jose; Keith van Rijsbergen; Mohan Kankanhalli und et al. (Herausgeber) *Proceedings of the 2014 International Conference on Multimedia Retrieval (ICMR)*, S. 411–414

[Shi94] Shi, Jianbo und Tomasi, Carlo: Good features to track, in: *Proceedings of the 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 593–600

[Shr13] Shree Devi, G.; Munir Ahamed Rabbani, M. und Jaya, A.: Face Recognition Using Principal Component Analysis with Median for Normalization on a Heterogeneous Data Set, in: *Proceedings of the 2013 International Conference on Information Technology Convergence and Services*, S. 147–153

[Sil03] Silva, Vin de und Tenenbaum, Joshua B.: Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems* (2003):S. 705–712

[Sim14] Simonyan, K. und Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556* (2014)

[Siv03] Sivic, Josef und Zisserman, Andrew: Video Google: a text retrieval approach to object matching in videos, in: *Proceedings of the 2003 International Conference on Computer Vision (ICCV)*, Bd. 2, S. 1470–1477

[Siv09] Sivalingam, Ravishankar; Morellas, Vassilios; Boley, Daniel und Papanikolopoulos, Nikolaos: Metric learning for semi-supervised clustering of Region Covariance Descriptors, in: *Proceedings of the 2009 ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, S. 1–8

[Sko84] Skovgaard, Lene T.: A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics* (1984), Bd. 11:S. 211–223

[Sme14] Smeulders, Arnold W. M.; Chu, Dung M.; Cucchiara, Rita; Calderara, Simone; Dehghan, Afshin und Shah, Mubarak: Visual Tracking: An Experimental Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2014), Bd. 36(7):S. 1442–1468

[Smi05] Smith, K.; Gatica-Perez, D.; Odobez, J. und Sileye, B.: Evaluating Multi-Object Tracking, in: *Proceedings of the 2005*

IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops, S. 36–43

[Sri16] Srinivas, Suraj; Sarvadevabhatla, Ravi Kiran; Mopuri, Konda Reddy; Prabhu, Nikita; Kruthiventi, Srinivas S. S. und Babu, R. Venkatesh: A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. *Frontiers in Robotics and AI* (2016), Bd. 2:S. 2654

[Sug16] Suggu, Sai Praneeth et al.: Hand in Glove: Deep Feature Fusion Network Architectures for Answer Quality Prediction in Community Question Answering, in: *Proceedings of the 2016 International Conference on Computational Linguistics (COLING) - Technical Papers*, S. 1429–1440

[Sze15] Szegedy, Christian; Wei Liu; Yangqing Jia; Sermanet, Pierre; Reed, Scott; Anguelov, Dragomir; Erhan, Dumitru; Vanhoucke, Vincent und Rabinovich, Andrew: Going deeper with convolutions, in: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–9

[Sze16] Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jon und Wojna, Zbigniew: Rethinking the Inception Architecture for Computer Vision, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2818–2826

[Tab14] Tabia, Hedi; Laga, Hamid; Picard, David und Gosselin, Philippe-Henri: Covariance Descriptors for 3D Shape Matching and Retrieval, in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 4185–4192

[Tan10] Tan, Xiaoyang und Triggs, Bill: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* (2010), Bd. 19(6):S. 1635–1650

[Ten00] Tenenbaum, Joshua B.; Silva, Vin de und Langford, John C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* (2000), Bd. 290(5500):S. 2319–2323

[Ten07] Tenenhaus, Arthur; Giron, Alain; Viennet, Emmanuel; Béra, Michel; Saporta, Gilbert und Fertil, Bernard: Kernel logistic PLS: A tool for supervised nonlinear dimensionality reduction and binary classification. *Computational Statistics and Data Analysis* (2007), Bd. 51(9):S. 4083–4100

[Teu15] Teutsch, Michael: *Moving Object Detection and Segmentation for Remote Aerial Video Surveillance: Zugl.: Karlsruhe, KIT, Diss., 2014*, Bd. 18 von *Karlsruher Schriften zur Anthropomatik*, KIT Scientific Publishing and Technische Informationsbibliothek u. Universitätsbibliothek, Karlsruhe and Hannover (2015), URL `http://edok01.tib.uni-hannover.de/edoks/e01fn15/82084845X.pdf`

[Tia15] Tian, Yonglong; Luo, Ping; Wang, Xiaogang und Tang, Xiaoou: Deep Learning Strong Parts for Pedestrian Detection, in: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, S. 1904–1912

[Tie04] Tieu, Kinh und Viola, Paul: Boosting Image Retrieval. *International Journal of Computer Vision* (2004), Bd. 56(1/2):S. 17–36

[Tom91] Tomasi, Carlo und Kanade, Takeo: *Detection and Tracking of Point Features*, Technical Report CMU-CS-91-132, Carnegie Mellon University (1991)

[Tor10] Torki, Marwan; Elgammal, Ahmed und Lee, Chan Su: Learning a Joint Manifold Representation from Multiple Data Sets, in: *Proceedings of the 2010 International Conference on Pattern Recognition (ICPR)*, S. 1068–1071

[Tos10] Tosato, D.; Farenzena, M.; Cristani, M. und Murino, V.: Part-based human detection on Riemannian manifolds, in: *Proceedings of the 2010 IEEE International Conference on Image Processing (ICIP)*, S. 3469–3472

[Tuz06] Tuzel, Oncel; Porikli, Fatih und Meer, Peter: Region Covariance: A Fast Descriptor for Detection and Classification, in: Aleš Leonardis; Horst Bischof und Axel Pinz (Herausgeber) *Proceedings of the 2006 European Conference on Computer Vision (ECCV)*,

Bd. 3952 von *Lecture Notes in Computer Science (LNCS)*, Springer Berlin Heidelberg (2006), S. 589–600

[Tuz07] Tuzel, Oncel; Porikli, Fatih und Meer, Peter: Human Detection via Classification on Riemannian Manifolds, in: *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8

[Tya08] Tyagi, Ambrish; Davis, James W. und Potamianos, Gerasimos: Steepest Descent For Efficient Covariance Tracking, in: *Proceedings of the 2008 IEEE Workshop on Motion and video Computing (WMVC)*, S. 1–6

[van07] van der Maaten, L.J.P.; Postma, E. O. und Herik, H. J. van den: Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research* (2007), (10-1)

[Var16] Varior, R. R.; Haloi, M. und Wang, G.: Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification, in: Bastian Leibe; Jiri Matas; Nicu Sebe und Max Welling (Herausgeber) *Computer vision - ECCV 2016*, Bd. 9909 von *Lecture Notes in Computer Science*, Springer, Cham (2016), S. 791–808

[Vem15] Vemulapalli, Raviteja und Jacobs, David W.: Riemannian Metric Learning for Symmetric Positive Definite Matrices. *arXiv:1501.02393* (2015)

[Vio01] Viola, P. und Jones, M.: Rapid object detection using a boosted cascade of simple features, in: *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. I–511 – I–518

[Vio03] Viola, Paul.; Jones, Michael J. und Snow, Daniel: Detecting pedestrians using patterns of motion and appearance, in: *Proceedings of the 2003 IEEE International Conference on Computer Vision (ICCV)*, Bd. 2, S. 734–741

[vL07] von Luxburg, Ulrike: A tutorial on spectral clustering. *Statistics and Computing* (2007), Bd. 17(4):S. 395–416

[Vu15] Vu, Tuan-Hung; Osokin, Anton und Laptev, Ivan: Context-Aware CNNs for Person Head Detection, in: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, S. 2893–2901

[Vur15] Vural, Elif und Guillemot, Christine: A study of the classification of low-dimensional data with supervised manifold learning. *arXiv:1507.05880* (2015)

[Wan05] Wang, Meng; Yang, Jie; Xu, Zhi-Jie und Chou, Kuo-Chen: SLLE for predicting membrane protein types. *Journal of Theoretical Biology* (2005), Bd. 232(1):S. 7–15

[Wan07] Wang, Xiaogang; Doretto, Gianfranco; Sebastian, Thomas; Rittscher, Jens und Tu, Peter: Shape and Appearance Context Modeling, in: *Proceedings of the 2007 IEEE International Conference on Computer Vision (ICCV)*, S. 1–8

[Wan09] Wang, Xiaoyu; Han, Tony X. und Yan, Shuicheng: An HOG-LBP human detector with partial occlusion handling, in: *Proceedings of the 2009 IEEE International Conference on Computer Vision (ICCV)*, S. 32–39

[Wan12] Wang, R.; Guo, H.; Davis, L. S. und Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification, in: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2496–2503

[Wan13] Wang, Xiaogang: Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters* (2013), Bd. 34(1):S. 3–19

[Wei04] Weinberger, K. Q. und Saul, L. K.: Unsupervised learning of image manifolds by semide .nite programming, in: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 988–995

[Wen11] Wengert, Christian; Douze, Matthijs und Jégou, Hervé: Bag-of-colors for improved image search, in: *Proceedings of the 2011 ACM International Conference on Multimedia*, S. 1437–1440

[Woj02] Wojtaszek, D. und Laganiere, R.: Using Color Histograms to Recognize People in Real Time Visual Surveillance, in: *Proceedings of the 2002 International Conference on Multimedia, Internet and Video Technologies*, S. 261–264

[Wu08] Wu, Yi; Wu, Bo; Liu, Jia und Lu, Hanqing: Probabilistic tracking on Riemannian manifolds, in: *Proceedings of the 2008 International Conference on Pattern Recognition (ICPR)*, S. 1–4

[Wu09] Wu, Yi; Jian Cheng; Jinqiao Wang und Hanqing Lu: Real-time visual tracking via Incremental Covariance Tensor Learning, in: *Proceedings of the 2009 IEEE International Conference on Computer Vision (ICCV)*, S. 1631–1638

[Wu12a] Wu, Yi; Cheng, Jian; Wang, Jinqiao; Lu, Hanqing; Wang, Jun; Ling, Haibin; Blasch, Erik und Bai, Li: Real-time probabilistic covariance tracking with efficient model update. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* (2012), Bd. 21(5):S. 2824–2837

[Wu12b] Wu, Yuwei; Ma, Bo und Li, Pei: A variational method for contour tracking via covariance matching. *Science China Information Sciences* (2012), Bd. 55(11):S. 2635–2645

[Wu15] Wu, Yuwei; Ma, Bo und Jia, Yunde: Differential tracking with a kernel-based region covariance descriptor. *Pattern Analysis and Applications* (2015), Bd. 18(1):S. 45–59

[Wu16] Wu, L. und Shen, C.: van den Hengel, A.: PersonNet: Person Re-identification with Deep Convolutional Neural Networks. *arXiv:1601.07255* (2016)

[Wu17] Wu, Lin; Chunhua Shen und van Hengel, Anton den: Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition* (2017), Bd. 65:S. 238–250

[Xi13] Xi, Li; Weiming, Hu; Chunhua, Shen; Zhongfei, Zhang; Dick, Anthony und van Hengel, Anton en: A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology* (2013), Bd. 4(4):S. 1–48

[Xia16] Xiao, Tong; Li, Hongsheng; Ouyang, Wanli und Wang, Xiaogang: Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1249–1258

[Yad08] Yadong Mu; Shuicheng Yan; Yi Liu; Huang, Thomas und Bingfeng Zhou: Discriminative local binary patterns for human detection in personal album, in: *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8

[Yan04] Yan, Ke und Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors, in: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 506–513

[Yan06] Yang, L. und Jin, R.: *Distance metric learning: A comprehensive survey*, Technical Report, Michigan State University, USA (2006)

[Yan14] Yang, Yang; Yang, Jimei; Yan, Junjie; Liao, Shengcai; Yi, Dong und Li, Stan Z.: Salient Color Names for Person Re-identification, in: David Fleet; Tomas Pajdla; Bernt Schiele und Tinne Tuytelaars (Herausgeber) *Computer Vision – ECCV 2014*, Bd. 8689 von *Lecture Notes in Computer Science*, Springer International Publishing, Cham (2014), S. 536–551

[Yao08] Yao, Jian und Odobez, Jean-Marc: Fast Human Detection from Videos Using Covariance Features, in: *Proceedings of the 2008 European Conference on Computer Vision (ECCV) - Visual Surveillance Workshop*

[Yao11] Yao, Jian und Odobez, Jean-Marc: Fast human detection from joint appearance and foreground feature subset covariances. *Computer Vision and Image Understanding* (2011), Bd. 115(10):S. 1414–1426

[Yi14] Yi, Dong; Lei, Zhen; Liao, Shengcai und Li, Stan Z.: Deep Metric Learning for Person Re-identification, in: *Proceedings of the 2014 International Conference on Pattern Recognition (ICPR)*, S. 34–39

[Yil06] Yilmaz, Alper; Javed, Omar und Shah, Mubarak: Object tracking: A survey. *ACM Computing Surveys* (2006), Bd. 38(4):S. 1–45

[Yoo06] Yoon, Kyongil; Harwood, David und Davis, Larry: Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation* (2006), Bd. 17(3):S. 605–622

[Zaj05] Zajdel, W.; Zivkovic, Z. und Krose, B.J.A.: Keeping Track of Humans: Have I Seen This Person Before?, in: *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA)*, S. 2081–2086

[Zha11a] Zhang, Junge; Huang, Kaiqi; Yu, Yinan und Tan, Tieniu: Boosted local structured HOG-LBP for object localization, in: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1393–1400

[Zha11b] Zhang, Ying und Li, Shutao: Gabor-LBP Based Region Covariance Descriptor for Person Re-identification, in: *Proceedings of the 2011 IEEE International Conference on Image and Graphics (ICIG)*, S. 368–371

[Zha13a] Zhang, Li Hong und Li, Lin: Improved Pedestrian Detection Based on Extended Histogram of Oriented Gradients. *Applied Mechanics and Materials* (2013), Bd. 347-350:S. 3815–3820

[Zha13b] Zhao, Rui; Ouyang, Wanli und Wang, Xiaogang: Unsupervised Salience Learning for Person Re-identification, in: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 3586–3593

[Zha16] Zhang, Ying; Li, Baohua; Lu, Huchuan; Irie, Atshushi und Ruan, Xiang: Sample-Specific SVM Learning for Person Re-identification, in: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1278–1287

[Zhe15] Zheng, Liang; Shen, Liyue; Tian, Lu; Wang, Shengjin; Wang, Jingdong und Tian, Qi: Scalable Person Re-identification: A Benchmark, in: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, S. 1116–1124

[Zhe16] Zheng, Liang; Yang, Yi und Hauptmann, Alexander: Person Re-identification: Past, Present and Future. *ArXiv:1610.02984* (2016)

[Zho17] Zhong, Zhun; Zheng, Liang; Cao, Donglin und Li, Shaozi: Re-ranking Person Re-identification with k-Reciprocal Encoding, in: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 3652–3661

[Zhu06] Zhu, Q..; Yeh, M.-C.; Cheng, K.-T. und Avidan, S.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients, in: *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1491–1498

Eigene Veröffentlichungen

[Hal15] Halbinger, Josef und Metzler, Jürgen: Video-Based Soccer Ball Detection in Difficult Situations, in: *Sports Science Research and Technology Support*, Bd. 464 von *Communications in Computer and Information Science*, Springer International Publishing (2015), S. 17–24

[Her11] Herrmann, Christian; Manger, Daniel und Metzler, Jürgen: Feature-based localization refinement of players in soccer using plausibility maps, in: *Proceedings of the 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, S. 672–678

[Her12] Herrmann, Christian; Metzler, Jürgen und Willersinn, Dieter: Semi-automatic people counting in aerial images of large crowds, in: *Proceedings of the 2012 SPIE Europe Security + Defence; Electro-Optical Remote Sensing, Photonic Technologies, and Applications V*, S. 85420Q

[Her13] Herrmann, Christian und Metzler, Jürgen: Density estimation in aerial images of large crowds for automatic people counting, in: *Proceedings of the 2013 SPIE Defense, Security, and Sensing; Airborne intelligence, surveillance, reconnaissance (ISR) systems and applications X*, S. 87130V

[Her15] Herrmann, Christian; Metzler, Jürgen; Willersinn, Dieter und Beyerer, Jürgen: Face- and appearance-based person

identification for forensic analysis of surveillance videos, in: *Proceedings of the 2015 Future Security*

[Her18] Herrmann, Christian; Metzler, Jürgen; Willersinn, Dieter und Beyerer, Jürgen: Distant pulse oximetry based on skin region extraction and multi-spectral measurement, in: *Proceedings of the 2018 SPIE Medical Imaging; Image-Guided Procedures, Robotic Interventions, and Modeling*

[Hüb08] Hübner, Yvonne; Metzler, Jürgen; Dürr, Bernhard; Jäger, Uwe und Willersinn, Dieter: Assessment and optimization of methods for tracking people in riot control scenarios, in: *Proceedings of the 2008 SPIE Europe Security + Defence; Electro-optical remote sensing, photonic technologies, and applications II*, Bd. 7114, S. 711404

[Jäg08] Jäger, Uwe; Höpken, Marc; Dürr, Bernhard; Metzler, Jürgen und Willersinn, Dieter: Multisensor benchmark data for riot control, in: *Proceedings of the 2008 SPIE Europe Security + Defence; Electro-optical remote sensing, photonic technologies, and applications II*, Bd. 7114, S. 711403

[Kro17] Kroschel, Kristian und Metzler, Jürgen: Berührungslose Bestimmung der Herz- und Atmungsfrequenz, in: *Berichtsband der Konferenz über elektronische Sprachsignalverarbeitung 2017*, Bd. 32

[Man14] Manger, Daniel und Metzler, Jürgen: Object detection in MOUT: evaluation of a hybrid approach for confirmation and rejection of object detection hypotheses, in: *Proceedings of the 2014 IS&T/SPIE Electronic Imaging*, S. 90240P

[Met07] Metzler, Jürgen und Willersinn, Dieter: Toward a sensor-based threat warning system for patrols in MOUT scenarios, in: *Proceedings of the 2007 SPIE Defense, Security and Sensing; Automatic Target Recognition XVII*, Bd. 6566, S. 65660T–1 – 65660T–9

[Met09] Metzler, Jürgen und Willersinn, Dieter: Robust tracking of people in crowds with covariance descriptors, in: *Proceedings of*

the 2009 SPIE Defense, Security, and Sensing; Visual Information Processing XVIII, Bd. 7341, S. 73410T

[Met10] Metzler, Jürgen und Willersinn, Dieter: Human detection in MOUT scenarios using covariance descriptors and supervised manifold learning, in: *Proceedings of the 2010 SPIE Defense, Security and Sensing; Visual Information Processing XIX*, Bd. 7701, S. 67942K–1 – 67942K–6

[Met12a] Metzler, Jurgen: Appearance-Based Re-identification of Humans in Low-Resolution Videos Using Means of Covariance Descriptors, in: *Proceedings of the 2012 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, S. 191–196

[Met12b] Metzler, Jürgen: Two-stage appearance-based re-identification of humans in low-resolution videos, in: *Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, S. 19–24

[Met13] Metzler, Jürgen und Pagel, Frank: 3D trajectory reconstruction of the soccer ball for single static camera systems, in: *Proceedings of the 2013 IAPR International Conference on Machine Vision Applications (MVA)*, S. 121–124

[Met14] Metzler, Jürgen; Monari, Eduardo und Kuntzsch, Colin: Application-driven merging and analysis of person trajectories for distributed smart camera networks, in: *Proceedings of the 2014 IS&T/SPIE Electronic Imaging*, S. 90260I

[Met15] Metzler, Jürgen: Context-based handover of persons in crowd and riot scenarios, in: *Proceedings of the 2015 IS&T/SPIE Electronic Imaging*, S. 94050Q

[Met17] Metzler, Jürgen; Kroschel, Kristian und Willersinn, Dieter: Automatic detection of measurement points for non-contact vibrometer-based diagnosis of cardiac arrhythmias, in: *Proceedings of the 2017 SPIE Medical Imaging*, S. 101351S

[Mül11] Müller, Thomas; Manger, Daniel und Metzler, Jürgen: Recognition of soccer players after occlusions using temporal color signatures, in: *Proceedings of the 2011 International Conference*

on Image Processing, Computer Vision, and Pattern Recognition (IPCV)

[Pag16] Pagel, Frank; Mossgraber, Jürgen; Tchouchenkov, Igor und Metzler, Jürgen et al.: A legally compliant multi-sensor system for security enhancement and real-time situation awareness in complex scenarios, in: *Proceedings of the 2016 Future Security*, S. 491–494

[Qu18] Qu, Chengchao; Metzler, Jürgen und Monari, Eduardo;: ivisX: an Integrated Video Investigation Suite for Forensic Applications, in: *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); Cross-Domain Biometric Recognition Workshop*

[Sch17] Schumann, Arne und Metzler, Jürgen: Person reidentification across aerial and ground-based cameras by deep feature fusion, in: *Proceedings of the 2017 SPIE Defense + Security; Automatic Target Recognition XXVII*, S. 102020A

Akronyme

aPKov *angepasster Porikli'scher Kovarianz-Tracker* 125

CRC *Crowd and Riot Control* 119

eKov *erweiterter Kovarianz-Tracker* 128

FHT *Farbhistogramm-Tracker* 132

FMT *farbbasierte Merkmals-Tracker* 131

HOG *Histogramme orientierter Gradienten* 20

ICF *Integral Channel Features* 24

ISM *Implicit Shape Models* 93

Isomap *Isometric Feature Mapping* 22

KCF *Kernelized Correlation Filters* 25

OTF *Objekt-Tracking-Fehler* 136

PCA-SIFT *Principal Component Analysis SIFT* 19

PV *positiver Vorhersagewert* 134

RPR *Richtig-Positiv-Rate* 134

RBF *radiale Basisfunktionen* 91

SDALF *Symmetry-Driven Accumulation of Local Features* 28

SIFT *Scale Invariant Feature Transform* 19

SVM *Support-Vektor-Maschinen* 29

TKFNN *tiefe künstliche faltende neuronale Netze* 15

VOT2015 *Visual Object Tracking VOT2015 Challenge* 40

VOT2016 *Visual Object Tracking VOT2016 Challenge* 40

Symbolverzeichnis

$\vec{\mathbf{0}}$	Nullvektor
$\vec{\mathbf{1}}$	Einsvektor
$\mathbb{R}^n$	n-dimensionaler euklidischer Raum

Kalligrafische Symbole

$\mathcal{A}, \mathcal{B}, \mathcal{T}$	Mengen von Kovarianzdeskriptoren
$\mathcal{L}$	Menge von Klassenlabels
$\mathcal{X}, \mathcal{Y}$	Mengen von Vektoren

Griechische Symbole

λ_i	Eigenwerte
$\boldsymbol{\mu}$	Mittelwertvektor von Merkmalsvektoren
$\boldsymbol{\Lambda}$	Eigenwertmatrix
$\boldsymbol{\Sigma}$	Kovarianzdeskriptor

$\boldsymbol{\Sigma_R}$ Kovarianzdeskriptor für einen Bildausschnitt $\boldsymbol{R}$

$\bar{\boldsymbol{\Sigma}}, \bar{\boldsymbol{\Sigma}}^*$ Mittelwerte im Sym_n^+

$\bar{\boldsymbol{\Sigma}}_{\log}$ logarithmierter $\bar{\boldsymbol{\Sigma}}$

$\overrightarrow{\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}}$ Tangentialvektor des Tangentialraums $T_{\boldsymbol{\Sigma}_0}\,\mathrm{Sym}_n^+$

$\boldsymbol{\Sigma} \sim (\bar{\boldsymbol{\Sigma}}, \boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}})$ normalverteilte Zufallsmatrix im Sym_n^+

$\Omega_{(\bar{\boldsymbol{\Sigma}}, \boldsymbol{Cov}_{\bar{\boldsymbol{\Sigma}}})}$ Mahalanobis-Distanz im Sym_n^+

Lateinische Symbole

b Breite eines Bilds oder Bildregion

d Dimension des niedrigdimensionalen euklidischen Zielraums

$\mathrm{d}(\cdot)$ euklidischer Abstand

e Dimension eines Merkmalsvektors

$\mathrm{g}(\cdot)$ geodätischer Abstand (riemannsche Metrik)

h Höhe eines Bilds oder Bildregion

$h(\cdot)$ Homöomorphismus

$\mathrm{k}(\cdot)$ Kosinus-Ähnlichkeit

$p(\cdot)$ Polynom

r_p, r_n, f_p, f_n Basismetriken

x, y Bildkoordinaten

x_i^h, y_i^h Koordinaten von Punkt-Hypothesen

x_i^s, y_i^s Koordinaten von Polygonschwerpunkten

$\boldsymbol{f}_{(x,y)}$ Merkmalsvektor

$\boldsymbol{c}_i, \boldsymbol{s}_i, \boldsymbol{t}_i$ vektorielle Repräsentation oberer Dreiecksmatrizen von $\boldsymbol{\Sigma}$

$\boldsymbol{v}_i$ Eigenvektoren

$\boldsymbol{x}$ Vektor des hochdimensionalen euklidischen Ursprungraums

$\boldsymbol{a}, \boldsymbol{b}$ Tangentialvektoren

$\boldsymbol{u}, \boldsymbol{y}$ Vektoren des niedrigdimensionalen euklidischen Zielraums

$L(\cdot)$ Längenfunktional

M Mannigfaltigkeit

D Dimension des hochdimensionalen euklidischen Ursprungraums

N_i Nachbarschaft von $\boldsymbol{x}_i$

Sym_n^+ riemannscher Raum der Kovarianzdeskriptoren

$T_{\boldsymbol{P}}M$ Tangentialraum am Punkt $\boldsymbol{P} \in M$

(M, g) riemannsche Mannigfaltigkeit

$\boldsymbol{A}$ symmetrische Adjazenzmatrix

$\boldsymbol{Cov}_{\tilde{\boldsymbol{\Sigma}}}$ Kovarianzmatrix normalverteilter Kovarianzdeskriptoren

$\boldsymbol{D}$ Knotengradmatrix (Diagonalmatrix)

$\boldsymbol{E}$ Einheitsmatrix

$\boldsymbol{F}[\cdot, \cdot]$ 2d-Array von Merkmalsvektoren

$\boldsymbol{P}[\cdot, \cdot, \cdot]$ 3d-Array von Integralbildern

$\boldsymbol{Q}[\cdot, \cdot, \cdot, \cdot]$ 4d-Array von Integralbildern

$\boldsymbol{I}_{\boldsymbol{R}}$ Ausschnitt $\boldsymbol{R}$ aus einem Intensitätsbild $\boldsymbol{I}$

$\boldsymbol{J_R}$ Ausschnitt $\boldsymbol{R}$ aus einem Integralbild $\boldsymbol{J}$

$\boldsymbol{I}_x, \boldsymbol{I}_y$ gefaltete Bildausschnitte

$\boldsymbol{L}$ Laplace-Matrix

$\boldsymbol{L'}$ normalisierte Laplace-Matrix

$\boldsymbol{P}$ Punkt einer Mannigfaltigkeit

$\boldsymbol{R}$ Bildregion (Bildausschnitt)

$\boldsymbol{S}^2$ 2-Sphäre

$\boldsymbol{T}, \boldsymbol{T}_a, \boldsymbol{T}_g$ Tracklet, Anfrage-Tracklet, Galerie-Tracklet

$\boldsymbol{V}$ Eigenvektormatrix

$\boldsymbol{W}$ gewichtete, symmetrische Adjazenzmatrix

$\boldsymbol{Y}$ Datenmatrix im euklidischen Zielraum