TECHNISCHE UNIVERSITÄT ILMENAU

Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Ilmenau

Face Recognition for Great Apes: Identification of Primates in Real-World Environments

Alexander Loos
geboren am 05.05.1983, Hildburghausen, Deutschland

Dissertation zur Erlangung des
akademischen Grades Doktor-Ingenieur (Dr.-Ing.)

Anfertigung im: Fachgebiet Elektronische Medientechnik
Institut für Medientechnik
Fakultät für Elektrotechnik und Informationstechnik

Gutachter: Prof. Dr.-Ing. Dr. rer. nat. h.c. mult. Karlheinz Brandenburg

Univ.-Prof. Dr.-Ing. habil. Gerhard Linß

Dr. Tilo Burghardt

Vorgelegt am: 07.05.2015

Verteidigt am: 13.01.2016

urn:nbn:de:gbv:ilm1-2016000034

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit am Fraunhofer-Institut für Digitale Medientechnologie in Ilmenau. Ich möchte an dieser Stelle dem Leiter des Instituts, Prof. Dr.-Ing. Karlheinz Brandenburg, für die Möglichkeit der Bearbeitung dieses Themas meinen Dank aussprechen. Ein besonderer Dank geht an Prof. Dr.-Ing. habil. Gerhard Linß und Dr. Tilo Burghardt für die Begutachtung meiner Arbeit. Ich danke außerdem Dr. Tobias Deschner für die Bild- und Videoaufnahmen, welche die Grundlage für eines der Vewendeten Datensets meiner Arbeit diente. Danke an den Zoo Leipzig und an das Wolfgang Köhler Primatenzentrum sowie an all die zahlreichen Assistenten und Tierpfleger für die Kollaboration während des SAISBECO Projektes. Ein herzliches Dankeschön geht auch an Laura Aporius und Karin Bahrke für das Aufnehmen und Annotieren von zahlreichen Trainings- und Testdaten. Vielen Dank auch an alle Mitglieder des SAISBECO Projektes, allen voran Andreas Ernst, Jens Garbas und Hjalmar Kühl für die kreative, unkomplizierte und fruchtbringende Zusammenarbeit.

Weiterhin möchte ich mich bei den zahlreichen Studenten bedanken, die maßgeblich zum Gelingen dieser Arbeit beigetragen haben. Ein großes Dankeschön geht außerdem an die Mitarbeiter und Mitarbeiterinnen sowie alle Doktoranden des Fraunhofer IDMT für das angenehme Arbeitsklima. Allen voran möchte ich mich bei meinen Kollegen der Gruppe Audio-visuelle Systeme, Uwe Kühhirt, Christian Weigel, Ronny Paduschek und Sascha Krämer für eure grenzenlose Geduld und eure wertvolle Unterstützung bei Implementierungsfragen bedanken. Ein herzliches Dankeschön geht auch an alle Mitglieder des SciBeCircles - durch unsere anregenden Meetings und fruchtbringenden Diskussionen habe ich nicht nur viel gelernt sondern konnte auch stets neue Anregungen für meine Arbeit gewinnen.

Des Weiteren möchte ich mich bei meinen Eltern, Hans-Joachim und Christa Loos, bedanken, ohne deren finanzielle Unterstützung mein Studium, welches die Grundlagen für diese Arbeit legte, nicht möglich gewesen wäre. Herzlichen Dank Anja und Alina, für Eure grenzenlose Geduld und eure Unterstützung in guten wie in schweren Zeiten während meiner Promotion. Nicht zuletzt geht mein Dank auch an alle anderen Freunde, Kommilitonen und Bekannte, die mir eine solch schöne und unvergessliche Zeit in Ilmenau bereitet haben.

Abstract

Due to the ongoing biodiversity crisis, many species including great apes such as chimpanzees or gorillas are threatened and need to be protected. To overcome the catastrophic decline of biodiversity, biologists recently started to use remote cameras for wildlife monitoring. However, the manual analysis of the resulting image and video material is extremely tedious, time consuming, and highly cost intensive.

To overcome the burden of exhaustive routine work, this thesis explores a novel application of computer vision: ***Automatic detection and identification of primates in visual footage***. Based on the assumption that humans and our closest relatives - the great apes - share similar properties of the face, algorithms for human face detection and recognition are adapted and extended in order to identify chimpanzees and gorillas in their natural habitats. The thesis proposes and evaluates an algorithmic framework comprising *detection*, *alignment*, and *identification* of primate faces in images and videos which is robust to various poses, challenging lighting, clutter, and partial occlusion among other extrinsic and intrinsic factors present in real-world environments. Free-living and captive chimpanzees and gorillas serve as sample species to demonstrate the feasibility but also highlight limitations of the approach.

Emphasizing on the identification part of the proposed system, first algorithms for reliable identification of primates in still images are proposed. Holistic global features and locally extracted descriptors are combined using a decision-level fusion paradigm. This approach is further extended to recognize great apes in videos. After faces have been detected and tracked through the sequence, a number of quality assessment modules are applied in order to select the frames which are best suited for subsequent recognition. Furthermore, a novel frame-weighting scheme is proposed which weights the predictions of multiple frames according to the classifier's confidence.

Secondly, the developed algorithms are thoroughly evaluated on realistic image and video benchmark datasets annotated by experts. To show the superiority of the proposed framework, it is compared to various state-of-the-art face recognition algorithms originally developed to identify humans.

In order to provide a practical proof of concept, the proposed algorithms are integrated into a unified full-automatic prototypical implementation currently used by biologists and ecologists to estimate population sizes faster and more precisely than current approaches. Thus, the proposed ***Primate Recognition Framework (PRF)*** has the potential to open up new venues in efficient wildlife monitoring and can help researchers to develop innovative protection schemes in the future.

Zusammenfassung

Aufgrund des gegenwärtigen Artensterbens sind viele Spezies, einschließlich Menschenaffen wie Schimpansen und Gorillas, vom Aussterben bedroht. Daher gewinnt die Überwachung der aktuellen Bestände mittels autonomer Aufnahmegeräte zunehmend an Bedeutung. Die manuelle Auswertung solcher Daten ist jedoch extrem mühsam, zeitaufwändig und kostenintensiv.

Um der immer größer werdenden Datenflut Herr zu werden, untersucht diese Arbeit ein neues Anwendungsgebiet der Bildverarbeitung und des maschinellen Sehens: ***Automatische Detektion und Identifikation von Primaten in Bildern und Videos***. Basierend auf der Annahme, dass Menschen und unsere nächsten Verwandten ähnliche Charakteristika des Gesichts aufweisen, werden in dieser Arbeit Algorithmen zur Erkennung menschlicher Gesichter erweitert um Schimpansen und Gorillas in ihrem natürlichen Lebensraum zuverlässig identifizieren zu können. Die Dissertation beschreibt und evaluiert ein algorithmisches System bestehend aus *Detektion*, *Ausrichtung* und *Identifikation* von Primatengesichtern in Bildern und Videos. Die vorgeschlagenen Algorithmen sind dabei robust gegenüber verschiedenen Posen, Beleuchtungsbedingungen und partiellen Verdeckungen sowie anderen Faktoren wie sie häufig in realen Anwendungsszenarien auftreten. Die Leistungsfähigkeit, aber auch Grenzen des Systems werden ausführlich anhand von Datensets freilebender und gefangener Schimpansen und Gorillas diskutiert.

Mit dem Fokus auf die individuelle Erkennung werden zuerst Algorithmen für eine zuverlässige Erkennung von Primaten in Bildern vorgestellt. Holistische Merkmale sowie lokale Deskriptoren werden mittels einer Entscheidungsfusion kombiniert. Anschließend wird dieser Ansatz auf die Erkennung von Menschenaffen in Videos erweitert. Nach der Detektion und Verfolgung von Gesichtern werden Module zur Qualitätsbeurteilung angewandt, um Frames zu identifizieren, die sich am besten für die folgenden Gesichtserkennungsalgorithmen eignen. Weiterhin wird ein neuartiger Frame-Weighting-Algorithmus beschrieben, welcher basierend auf der Konfidenz des Klassifikators die Resultate mehrerer Frames gewichtet. Des Weiteren werden die entwickelten Algorithmen auf realistischen, von Experten annotierten Bild- und Videodatenbanken, sorgfältig evaluiert. Um die Vorteile des vorgeschlagenen Systems zu demonstrieren, wird es mit anderen dem Stand der Technik entnommenen Algorithmen zur Gesichtserkennung verglichen.

Die implementierten Algorithmen wurden in einer prototypischen Anwendung zusammengeführt, welche derzeit von Biologen genutzt wird um Populationsgrößen schneller und genauer schätzen zu können. Daher hat das entwickelte ***Primate Recognition Framework (PRF)*** das Potential, den Weg zu effizienteren Monitoringverfahren zu ebnen und damit zukünftig Wissenschaftlern zu helfen, neue innovative Schutzmaßnahmen zu entwickeln.

Contents

1. Introduction

1.1. Motivation

Due to the ongoing biodiversity crisis, many species are on the brink of extinction. According to the *International Union for Conservation of Nature (IUCN)* about 22% of the mammal species worldwide are threatened or extinct [13]. The current biodiversity crisis is observed all over the world. Primates such as chimpanzees or gorillas are hit by the crisis and belong to a species that is severely endangered. Walsh *et al.* [14] for instance reported a decrease of ape populations in western equatorial Africa by more than a half between 1983 and 2000. Similar conclusions were drawn by Campbell *et al.* in [15]. They observed a 90% decrease of chimpanzee sleeping nests in Côte d'Ivoire, West Africa, between 1990 and 2007.

Those agitating results demonstrate the urgent need to intensify close surveillance of this threatened species in order to protect the remaining populations. Many protective areas have already been established. However, effectively protecting animals requires good knowledge of existing populations and fluctuations of population sizes over time. Individual identification of animals is not only a prerequisite to measure the success of implemented protection schemes but also for many other biological questions such as wildlife epidemiology, social network analysis, and behavioral ecological research. However, it is a labor intensive task to estimate population sizes in the wild. A widely used population monitoring approach for instance approximates species density by counting animal traces such as nests. This information is subsequently converted into individual abundance [16, 17]. To overcome the burden of such tedious tasks, non-invasive monitoring techniques which are based on automatic camera traps are currently under development and the number of published studies that utilize autonomous recording devices is tremendously increasing [18]. However, the collected data is mostly analyzed manually which is a time and resource consuming task. Consequently, there is a high demand for automated algorithms which are able to assist biologists in their effort to analyze remotely gathered video recordings. Especially so-called capture-mark-recapture methods, commonly used in ecology, could benefit from an automated system for animal identification.

This thesis presents a completely autonomous framework for individual identification of free-living as well as captured primates in images and videos. Based on the assumption that humans and our closest relatives, the great apes, share similar properties of the face, the proposed Primate Recognition Framework (PRF) is built upon face detection and face recognition technology.

Due to its close to real-time capability, a second novel applicability of the proposed PRF is within the *edutainment domain*, where the system might be deployed as an interactive tool for individual identification of great apes in zoos or wildlife parks. The developed system for non-intrusive wildlife observation thus might give visitors the chance to apply innovative technology in a novel real-world scenario. More importantly, it could open up an opportunity for the wider public to obtain deeper insights into the social complexity of chimpanzee or gorilla groups. Thus, a concrete application in the field of ecology and behavioral research could be established which might consequently lead to increasing environmental awareness and species protection. Moreover, the developed system for individual identification of primates might also be applied in the field of *citizen science*. With the help of zoo visitors new datasets could be established which might help biologists to obtain new insights into the social behavior and communication complexity of primates which currently is a large and growing field of biological research [19, 20, 21].

1.2. Wildlife Monitoring: A Brief Introduction

To fully understand the practical implications of the proposed PRF, a brief review of existing wildlife population monitoring techniques is given in this section.

According to [22], three main approaches for estimating population parameters such as density or fluctuation rates are commonly used within the biological community. The first approach is referred to as *occupancy method*. Here, the density and population size as well as the distribution of species is estimated by repeatedly confirming species presence or absence in different sampling locations [23]. The second technique approximates species density and population sizes by counting animal traces such as nests. This information is subsequently used to estimate individual abundance [16, 17]. However, this approach is not only tedious and time consuming but also prone to errors since it requires auxiliary parameters based on statistical assumptions which are hardly to be met in real-world situations [22]. The third approach, which nowadays is commonly used by biologists and ecologists, refers to so-called *capture-mark-recapture* techniques [24, 25, 26]. A sample of individuals of a population is identified in the first phase, i.e. individuals are *capture*d and *mark*ed. In the next step a second group of individuals is identified. Some of the individuals captured and marked in the first phase of the approach are also present in the second group, i.e. they are *recapture*d. This information can then be exploited to estimate the population size using statistical approaches. The latter method, although extensively used by biologists, requires reliable identification and recognition of individuals which can be tedious in wildlife environments. Early approaches for individual identification concentrated on intrusive analysis in which animals are physically marked by ear tags, leg bands or dyes [27, 28].

However, a number of studies showed that such methods can cause extreme discomfort and pain. Furthermore, invasive methods are also known to be a severe intrusion into ecosystems itself and may therefore bias observation due to potential changes in behavior, reproduction, and even survival [29, 30, 31]. Therefore, non-invasive monitoring techniques using autonomous video recording devices have been used extensively by biologists and ecologists in recent years [18, 32]. In a study by Head *et al.* [33] it has been proven successful to place video recording devices with a two-fold strategy: First, a systematic grid with cell size of 1×1 km is placed across the study area. Secondly, within each cell, cameras are placed at locations that are frequently visited by wildlife such as animal trails, tree bridges across rivers or swamps, tool use sites or important feeding spots such as fruit trees. Figure 1.1 shows examples of camera traps often used by biologists. Furthermore, typical placement sites to monitor great ape species in national parks and protected areas are illustrated. Screenshots of captured videos using such remote recording devices can be seen in Figure 1.2

(a) (b) (c)

Figure 1.1.: ***Examples of autonomous recording devices.*** The figure shows examples of infrared-triggered autonomous camera traps (a) used by biologists and ecologists to monitor great ape populations in their natural habitats. Typical locations for placing autonomous video cameras are either honey extraction sites (b) or animal trails (c). Images source: [22, 1]. [I01]

However, for most practical applications the important information - namely the number of individuals present in the video - often has to be extracted manually which is not only tedious, time consuming, and cost intensive but also prone to errors due to subjectivity and fatigue. Hence, innovative automatic procedures are required to overcome the burden of tedious and time consuming routine work. Thus, previously bound human resources could then be released to conduct innovative and important research.

Figure 1.2.: ***Screenshots of video footage captured from autonomous recording devices.*** The figure shows screenshots of video sequences captured by remote video cameras at Loango National Park, Gabon, Africa (top row: chimpanzees, second row: gorillas). A large variety of different lighting conditions, head poses, facial expressions, and limited video resolution are present in most video recordings which place high demands on a facial recognition system. [I02,I03]

1.3. Problem Outline and Contributions

This thesis proposes and evaluates a unified automatic framework for the detection and identification of great apes in images and videos recorded in real-world environments. As stated earlier, there is an urgent need for automatic routine procedures to detect, identify and recognize animals filmed by autonomous recording devices in their natural habitats. The emerging and interdisciplinary field of research commonly known as *Animal Biometrics* has recently attracted the attention of many ecologists, biologist and computer vision experts. As will be shown in this thesis, a large and growing amount of literature has recently been published in which full- or semi-automatic algorithms to detect and recognize animals in audio-visual footage were proposed. Most of these techniques utilize unique markings known as coat patterns on fur or skin of animal bodies as key features for automatic identification. However, many species - including great apes - either do not carry any obvious coat patterns or such markings cannot be efficiently exploited for identification due to insufficient image or video resolution.

Based on the assumption that humans and our closest relatives share similar properties of the face, ideas originating for human face detection and recognition technology are adapted and extended in this thesis to non-intrusively identify primates in natural environments. Although the proposed Primate Recognition Framework (PRF) is theoretically applicable to all four great apes species, particular emphasis is placed on chimpanzees and gorillas as two representative primate species.

The main purpose of the system is to support biologists, ecologists, and gamekeepers in annotating huge amounts of video footage gathered in natural habitats of great apes. Thus, high demands are placed on the system in terms of various uncontrollable factors comprising ***lighting*** conditions, ***non-cooperative*** subjects, a variety of different ***facial expressions***, ***pose*** variations, limited ***resolution*** of video recordings, and even partial ***occlusion*** by branches, leafs, and conspecifics just to name a few. In addition to robustly identifying primate genuine individuals, a further requirement of the proposed PRF is to reliably reject unknown individuals that have not yet been included into the training set. Once enough high-quality material of a previously unknown individual has been gathered, it should be possible for biologists to retrain the system in an intuitive and user-friendly fashion.

As claimed earlier, a second application scenario of the proposed PRF might be within the edutainment domain. The face detection and identification framework can be used by zoo visitors to monitor captured primate individuals and get immediate feedback about the identity of the observed individual. Since in this scenario the number of individuals is known a-priori, rejection of unknown individuals is not required. However, the system has to be close to real-time capable in order to give immediate feedback to the user.

The primary goal of this thesis is to provide a proof of concept, and thus demonstrate that face detection and recognition algorithms, originally developed to identify humans, can be extended to reliably recognize great apes in real-world environments. This task involves explanation and justification of theoretical concepts such as feature extraction, feature space transformation, and classification with a special focus on high robustness against above mentioned challenges in natural environments. Moreover, these techniques are efficiently implemented to provide a prototypical close to real-time capable tool for biologists, ecologists, and gamekeepers. Finally, the proposed PRF is thoroughly evaluated on realistic datasets of captive as well as free-living chimpanzee and gorilla individuals gathered under real-world conditions. The developed system is compared to other state-of-the-art face recognition algorithms to show the superiority of the proposed algorithms for the application presented in this thesis. To summarize, the contributions of this thesis are as follows:

1. An extensive and thorough **literature review** of algorithms in the field of *Visual Animal Biometrics* is given. Available state-of-the-art software tools and algorithms to automatically detect, represent, and interpret phenotypic appearance of various animal species in images and videos are presented and discussed. Advantages and shortcomings of different approaches are reviewed which to the best of the author's knowledge has not yet been done to this extend.

2. A detailed description of the proposed **Primate Recognition Framework (PRF)** is given. First, a robust facial descriptor referred to as Enhanced Local Gabor Ternary Pattern Histogram Sequence (ELGTPHS) is proposed which efficiently encodes the visual traits of a primate's face. Secondly, an effective processing chain comprising state-of-the-art feature space transformation and classification techniques is suggested for fast and accurate identification. Thirdly, the robustness of the proposed system against various intrinsic and extrinsic factors is further enhanced by combining the results of global and local features in a decision-level based manner for accurate recognition.

3. The developed framework for primate photo identification is extended to accurately identify great ape individuals in **video sequences**. Once faces of chimpanzees or gorillas have been detected they need to be tracked through the video. To further enhance the accuracy of the system and at the same time maintain near real-time performance, the extracted face-track is subsequently scanned by quality-assessment routines to select the frames which are best suited for recognition. Moreover, a novel frame-weighting procedure is proposed which combines the results of multiple frames by taking classification confidences into account.

4. All proposed algorithms are integrated into a unified full-automatic **prototypical implementation** currently used by biologists and ecologists for wildlife monitoring, population counting, and species protection. The developed prototype is not only capable of identifying individual members of a population but can also be configured to reject unknown individuals. Thus, once a reasonable amount of annotated facial images of previously unseen animals is available, the proposed PRF can be retrained by biologists in a user-friendly way in order to recognize new individuals.

5. As a proof of concept, a thorough **evaluation** of the proposed system on realistic datasets of different captive and free-living individuals gathered in natural environments is conducted. The developed PRF is benchmarked against state-of-the-art human face recognition algorithms in order to proof the superiority of the proposed identification system for the task at hand. It is shown that additional representative training data - real or synthesized data rendered from a generic 3D model - can be used to further enhance the robustness of the system against various difficulties present in real-world situations.

1.4. Thesis Overview

The thesis is organized into seven chapters, where Chapter 4 represents the core of the thesis where the proposed Primate Recognition Framework (PRF) is described in detail.

Chapter 2 introduces the background considered to be relevant to understand the subsequent chapters. Particular emphasis is placed on fundamentals of pattern recognition including techniques and algorithms for feature extraction, feature space transformation, and classification.

Chapter 3 reviews the state-of-the-art of the research fields of interest. First, an extensive and thorough literature survey of the new and emerging field known as *Visual Animal Biometrics* is given. Techniques and algorithms for automatic animal detection, species classification, as well as individual identification are categorized and reviewed. Advantages and shortcomings of approaches proposed in the recent past are discussed and highlighted with respect to the objectives of this thesis. A compact yet comprehensive review of existing biometric identification systems based on facial analysis is reviewed in the second part of this chapter. Particular emphasis is placed on holistic techniques which are considered to be applicable for the task at hand. Some of the surveyed algorithms for face recognition of human beings are compared with the proposed framework in Chapter 5.

Chapter 4 is the core chapter of this thesis and thus introduces the proposed framework. First, algorithms are proposed which are applicable for still images only. A robust global face descriptor called ELGTPHS is proposed which unites the strengths of two descriptors, Gabor-based features and Extended Local Ternary Patterns (ELTP). It is then shown how a sophisticated feature space transformation technique known as Locality Preserving Projections (LPP) can be used to project the obtained high-dimensional feature vectors into a smaller yet more disjunctive subspace. A recently proposed algorithm called Sparse Representation Classification (SRC) which is based on Compressed Sensing (CS) theory, a mathematical framework for efficient signal measurement and reconstruction, is used for classification. Based on the assumption that different features tend to misclassify different patterns, a decision fusion scheme is proposed which merges the results of global and local features to further enhance the accuracy and robustness of the system. In the second part of this chapter, the proposed algorithms for facial feature analysis of primates are extended to robustly identify African great apes in videos. A prerequisite for accurate identification in videos is robust detection and tracking of faces through the sequence. The resulting face-tracks are subsequently analyzed in order to select the frames which are best suited for subsequent recognition. A number of different quality-assessment modules including algorithms for pose and visual quality estimation are proposed for that purpose. Finally, a novel frame weighting scheme is introduced which is based on Maximum Likelihood Estimation (MLE) and Bayesian statistics to obtain a final decision.

In **Chapter 5**, the proposed PRF is thoroughly evaluated on real-world datasets of captive and free-living chimpanzees and gorillas. Detailed information about all data used for experimentation is given, followed by an introduction of the applied evaluation measures. Experiments are first conducted in a closed-set fashion where the number of individuals in a dataset is known a-priori. The developed identification system is benchmarked against other state-of-the-art face recognition algorithms which were originally developed to identify humans. Various pre-processing steps for face alignment and lighting normalization are thoroughly investigated to identify the techniques which are best suited for the application presented in this thesis. It is further shown that additional data - real or artificial data synthesized from a generic 3D model of a chimpanzee head - can be used to enhance the system's robustness against challenges frequently present in real-world environments. Secondly, experiments are conducted in an open-set fashion to show that the proposed system is not only capable of reliably identifying known individuals but also rejecting subjects which are not yet included in the training set. Moreover, the developed framework is also tested on annotated video sequences. The proposed frame-weighting scheme is compared to other video-based face recognition approaches commonly used within the research community.

Chapter 6 presents the prototypical implementation of the proposed algorithms currently used by biologists of the Max-Planck-Institute as supportive tool for annotating huge amounts of video footage gathered in protected areas and national parks. The real-world prototype of the proposed PRF consists of three main parts combined into a single unified framework: The *Training Module* offers the possibility to add new training data of known individuals or previously unseen subjects to build an extended model which can subsequently be used for identification. The *Ripper Module* provides the core functionality of the developed framework. Videos gathered in natural habitats can be processed in order to detect and identify primate individuals present in the video. The extracted information is written into an SQL database which can be efficiently searched by biologists by applying common SQL commands. Finally, a video sequence and the according SQL database can be loaded with the *Graphical Interface*. The obtained results, i.e. information about the location and names of individuals as well as obtained confidence measures, are displayed. A separate timeline is created for each individual present in the video which allows efficient browsing through video sequences.

Chapter 7 concludes and summarizes the thesis. Achievements and limitations of the proposed framework are discussed and future directions of research which might further enhance the developed methods for non-intrusive identification of great apes are indicated.

2. Background

2.1. Chapter Overview

This chapter gives an introduction to the fundamentals of pattern recognition considered relevant for understanding the theoretical and practical aspects of the proposed Primate Recognition Framework (PRF). The overall system comprises three main components: *Detection*, *Alignment*, and *Identification*. However, the emphasis of this thesis lies on the identification of great apes in their natural habitats, assuming that faces of primates have already been detected and tracked through the video sequence. In fact, a face detection library named Sophisticated High-Speed Object Recognition Engine (SHORE™), originally developed to detect human faces, is utilized for that purpose. Details about the detection of primate faces are given in Section 4.2.1.

The core functionality of PRF - the identification of great apes - follows a standard pattern recognition pipeline: *Feature Extraction*, *Feature Space Transformation*, and *Classification*. Thus, the fundamentals of those three problems are covered in more detail in this section, emphasizing techniques either used within the proposed PRF directly or in state-of-the-art face recognition algorithms used for benchmarking purposes.

Section 2.2 gives an introduction to visual feature extraction algorithms covering global features as well as local keypoint descriptors. For many pattern recognition problems, especially for face recognition, feature space transformation techniques are indispensable to perform robust and fast recognition in practice. Therefore, three of the most common feature reduction techniques are discussed in **Section 2.3**. Finally, **Section 2.4** briefly introduces both classical and more recently developed classification paradigms commonly utilized in many face recognition applications.

2.2. Visual Descriptors

This section gives a brief summary of some important state-of-the-art visual feature extraction algorithms and techniques. A *feature* can be seen as an abstract, often application specific property of the content of images or videos. The aim of feature extraction is to find a compact yet meaningful representation of these features which leads to a so called visual *descriptor*. Low-level descriptors can represent a number of different features including texture, color, shape, motion, etc.

Hence, descriptors are one key step to characterize the connection between spatially and temporally discretized functions of real world scenes in form of pixels and what humans recall after having observed an image or a sequence of images. A suitable representation for visual features therefore enables a number of automatic procedures including individual identification, object categorization, image region classification, image segmentation and many more.

Since this field of research has been very active within the last decade, a vast number of types, methods, and applications exist, making a taxonomic classification of algorithms difficult. For the sake of simplicity, visual features are therefore grouped into two categories within this thesis: *global* and *local* descriptors. Different definitions of global and local features exist in the literature. Throughout this thesis the following definition is used: While global descriptors give a representation of the whole image, local descriptors encode the information of a restricted region around an interest point. This section particularly emphasizes the global and local feature descriptors used in this thesis. Since the overall objective of the proposed PRF is to facilitate recognition of great apes in visual footage independent from the type of recording (e.g color, gray-scale, infrared), this section solely reviews visual texture descriptors. For a more thorough review of visual descriptors the interested reader is referred to [34, 35, 36].

2.2.1. Global Features

Some of the most well established texture descriptors for pattern recognition are Gabor-based features [37], Local Binary Patterns (LBP) [38, 39], and its extension Local Ternary Patterns (LTP) [40]. Since those features represent a key aspect for the proposed PRF they are reviewed within the subsequent sections.

Gabor Features

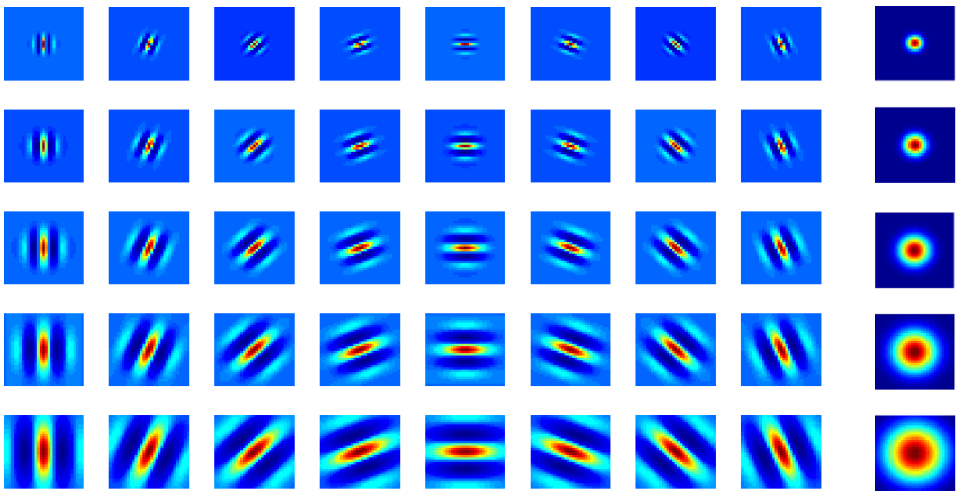
Gabor functions were first proposed by Dennis Gabor in 1946 as a complex sinusoidal wave multiplied with a Gaussian function [37]. In 1985 it was discovered that simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions. Thus, image analysis by Gabor wavelets provides similarity to perception in the human visual system [41].

Complex two-dimensional Gabor kernels at pixel location $z = (x, y)$ are defined as

$$\mathbf{\Psi}_{\mu,\nu}(z) = \frac{|k_{\mu,\nu}|^2}{\sigma^2} e^{\frac{-|k_{\mu,\nu}|^2|z|^2}{2\sigma^2}} [e^{ik_{\mu,\nu}z} - e^{\frac{-\sigma^2}{2}}], \tag{2.1}$$

where $k_{\mu,\nu} = k_\nu e^{i\theta_\mu}$ with $k_\nu = \frac{k_{max}}{f^\nu}$ and $\theta_\mu = \frac{\pi\mu}{F}$. The maximum frequency is denoted as k_{max}, f is the spacing between kernels in the frequency domain, and F is the number of rotations defined by the user. Furthermore, σ represents the ratio of the Gaussian window to the wavelength.

Every kernel is the product of a Gaussian envelope and a sinusoidal plane wave. Figure 2.1 shows the real part of Gabor kernels at five different scales $\nu \in \{0, \ldots, 4\}$ and eight orientations $\mu \in \{0, \ldots, 7\}$ (a) as well as their magnitudes (b). The kernels exhibit desirable characteristics of spatial frequency, spatial locality, and orientation selectivity. A more detailed introduction to 2D Gabor wavelets can be found in [42].



(a) (b)

Figure 2.1.: ***Gabor kernels for five different scales and eight orientations as well as their magnitudes.*** The figure shows the real part of Gabor kernels at five scales $\nu \in \{0, \ldots, 4\}$ and eight orientations $\mu \in \{0, \ldots, 7\}$ (a) as well as their magnitudes (b). For the generation of Gabor kernels the parameters were set as follows: $k_{max} = \frac{\pi}{2}$, $f = \sqrt{2}$, $\sigma = \pi$. The kernels exhibit desirable characteristics of spatial frequency, spatial locality, and orientation selectivity.

The Gabor wavelet transform is defined as the convolution of a grayscale input image $\boldsymbol{I}(z)$ with a set of two-dimensional Gabor kernels $\boldsymbol{\Psi}_{\mu,\nu}(z)$.

$$\boldsymbol{G}_{\mu,\nu}(z) = \boldsymbol{I}(z) * \boldsymbol{\Psi}_{\mu,\nu}(z), \tag{2.2}$$

where $\boldsymbol{G}_{\mu,\nu}(z)$ is the output image for orientation μ and scale ν at pixel location $z = (x, y)$ and $*$ denotes the convolution operator. Due to the multi-resolution and multi-orientation properties of the Gabor wavelet transform it provides a robust measurement of the local spectral energy density concentrated around a given position and frequency in multiple directions. Due to its desirable properties, the Gabor wavelet transform has been used in many image analysis applications such as texture segmentation [43], edge detection [44], and face recognition [45, 46, 47, 48].

In general, $\boldsymbol{G}_{\mu,\nu}(z)$ is complex and can be rewritten as $\boldsymbol{G}_{\mu,\nu}(z) = |\boldsymbol{G}_{\mu,\nu}(z)|e^{i\boldsymbol{\Phi}_{\mu,\nu}(z)}$, where $|\boldsymbol{G}_{\mu,\nu}(z)|$ denotes the Gabor Magnitude Picture (GMP) and $\boldsymbol{\Phi}_{\mu,\nu}(z)$ is the phase at pixel location z, scale ν, and orientation μ. Although it was shown in [47] that $\boldsymbol{\Phi}_{\mu,\nu}(z)$ can contain discriminative information, often only $|\boldsymbol{G}_{\mu,\nu}(z)|$ is used for further processing since it contains the local energy variation in a given image which is the most important cue for robust and accurate recognition [42]. In many applications the GMPs are used directly as descriptors. In this thesis however the generation of the GMPs only serves as a pre-processing step to extract the actual face descriptor (see section 4.2.3.1 for details).

Local Binary Patterns (LBP)

The main idea of Local Binary Patterns (LBP) is to model the local texture of an image by using the joint distributions of differences between a center pixel and its surrounding neighbors. The original LBP descriptor as proposed by Ojala *et al.* in [38] thresholds each 3×3 block within a monochrome image based on its center pixel to obtain a binary number for each block. Figure 2.2 shows an illustration of the basic LBP operator applied on a chimpanzee face.

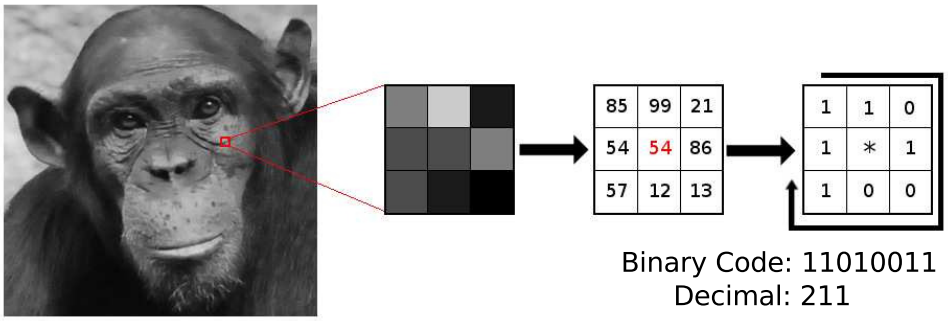


Figure 2.2.: ***The basic LBP operator.*** The figure illustrates the basic LBP operator as proposed in [38]. The eight neighbor pixels are thresholded based on the gray value of the center pixel. A binary number is obtained by assigning a "1" to pixels with equal or higher gray values and a "0" to sampling points with values lower than the center pixel. For eight neighbors, $2^8 = 256$ different labels can be obtained.

As the neighborhood of each pixel consists of 8 neighbors, $2^8 = 256$ different labels can be obtained depending on the gray values of the center pixel and its eight neighbors. The statistics of these labels in form of histograms is then commonly used to model the local texture of images.

The original LBP descriptor was later presented in a more general form by the same authors in [39]. Let $\boldsymbol{I}$ be a gray-scale image where $z_c = \boldsymbol{I}(x_c, y_c)$ denotes the gray level value of a pixel at location (x_c, y_c). Moreover, $z_p = \boldsymbol{I}(x_p, y_p)$ represents the gray-level value of a pixel within an evenly spaced circular neighborhood of P sampling points $p = 0, \ldots, P$ and radius R around

(x_c, y_c) with

$$x_p = x_c + R\cos\left(\frac{2\pi p}{P}\right) \quad \text{and} \tag{2.3}$$

$$y_p = y_c - R\sin\left(\frac{2\pi p}{P}\right). \tag{2.4}$$

Figure 2.3 shows three example patterns of different circular neighborhoods and sampling points. Bilinear interpolation is applied in case the location of $\boldsymbol{I}(x_p, y_p)$ is not in the center of a pixel.

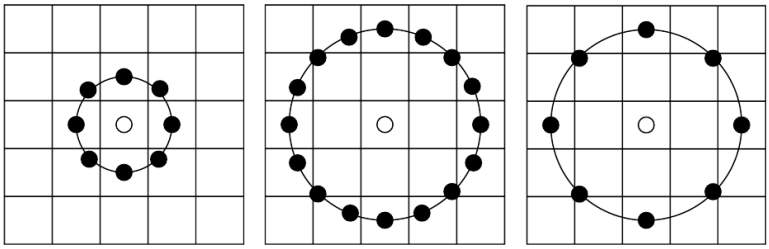


Figure 2.3.: ***Circular neighbors for different (P,R).*** The figure shows the circular neighbors obtained from equation 2.3 for (8,1) (a), (16,2) (b), (8,2) (c). Bilinear interpolation is applied for sampling points that are not located at the center of a pixel in order to obtain the gray-level values of these points. Image source: [49]

The generic LBP operator for radius R and number of sampling points P is then given by

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(z_p - z_c)2^p, \tag{2.5}$$

where $s(z_p - z_c)$ is a thresholding function defined as

$$s(z_p - z_c) = \begin{cases} 1, & z_p - z_c \geq 0 \\ 0, & z_p - z_c < 0. \end{cases} \tag{2.6}$$

Hence, the signs of the differences between a center pixel and its neighbors are interpreted as a P-bit binary number resulting in 2^P possible values.

In practice, the local gray value distribution is often described using a histogram with 2^P bins. Figure 2.4 shows a gray scaled image of a chimpanzee face, the corresponding LBP image, and the according LBP histogram for $R = 1$ and $P = 8$.

(a) (b) (c)

Figure 2.4.: ***LBP image and according histogram.*** The figure shows an input image (a), the corresponding LBP image (b), and the according LBP histogram (c) for radius $R = 1$ and number of sampling points $P = 8$.

Note that the generic $\text{LBP}_{8,1}$ operator is very similar to the basic LBP proposed in [38]. However, an important difference of the generic LBP descriptor is that the neighborhood is defined circularly which generally results in sampling points that are not located in the center of a pixel. The gray-level values of such pixels are usually defined by bilinear interpolation.

Ojala *et al.* presented another extension of the basic LBP operator in [39] which utilizes so called *uniform patterns*. A uniformity measure U was introduced which denotes the number of bitwise transitions from "0" to "1" or vice versa when the binary pattern is considered circular. A pattern is called uniform if U is at most two. For instance, the binary patterns 00111100 (2 transitions) and 11111111 (0 transitions) are uniform while 00110011 (4 transitions) and 01010100 (6 transitions) are not. Hence, uniform patterns can be seen as a representation of primitive structural information such as edges, spots, flat areas or corners [49]. The uniform LBP operator is denoted as $\text{LBP}^{u2}_{P,R}$ for the remainder of the thesis. When mapping uniform LBP patterns into a histogram, each uniform binary code is assigned to a separate bin and there is only one single label for all non-uniform patterns. Thus, the number of possible labels for a P-bit binary number is reduced from 2^P to $P(P-1)+3$ [49]. There are two major advantages of using uniform LBP patterns: First, for real-world images most binary patterns are in fact uniform. Ahonen *et al.* for instance found in [50] that 90.6% of $\text{LBP}_{8,1}$ patterns and 85.2% of LBP patterns in the (8,2) neighborhood are uniform for images of a human face. Consequently, non-uniform $\text{LBP}_{P,R}$ patterns often lead to sparse histograms, i.e. feature vectors with lots of zero-valued entries. Omitting non-uniform patterns drastically reduces the dimensionality of the final feature vector which in turn leads to a more dense representation of the texture. Secondly, it was shown that uniform LBP features lead to more robust and accurate recognition [39, 50, 49]. This indicates that by considering uniform patterns, texture descriptors can be obtained which are are less prone to noise and therefore more reliable under natural conditions than their non-uniform counterparts.

Local Ternary Patterns (LTP)

One known disadvantage of LBP, however, is that it may not work well for noisy images or flat regions such as cheeks or the forehead of human faces due to its thresholding paradigm which is solely based on the gray level value of the center pixel. Moreover, the reliability of LBP is known to decrease significantly for large illumination changes and shadowing. To overcome these limitations, Tan and Triggs proposed to replace LBP with a three-level operator called Local Ternary Patterns (LTP) in [40]. In LTP the difference of the center pixel and its surrounding neighbors is encoded using three values (-1,0,1) according to a user specified threshold t. Thus, the thresholding function $s(n)$ of equation 2.6 is replaced with $s(z_p, z_c, t)$ defined as

$$s(z_p, z_c, t) = \begin{cases} 1, & z_p \geq z_c + t \\ 0, & |z_p - z_c| < t \\ -1, & z_p \leq z_c - t. \end{cases} \tag{2.7}$$

Figure 2.5 illustrates the process of applying the basic LTP operator on a chimpanzee face where the threshold t was set to 5. Since the center pixel has a gray value of 54, the tolerance interval is thus $(49, 59)$.

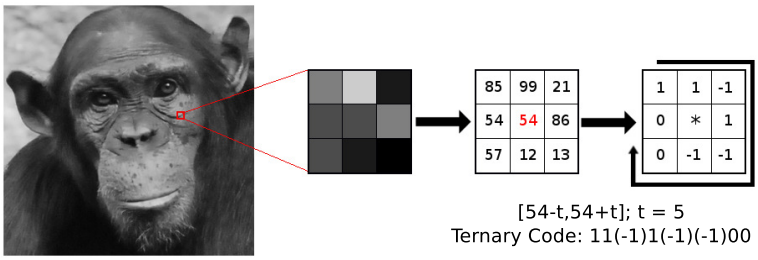


Figure 2.5.: ***The basic LTP operator.*** The figure shows how LTPs are obtained from an input image. In this example the threshold t is set to 5. Hence, the tolerance interval is $(49, 59)$. Neighboring pixels with gray values that fall within this interval are encoded as "0" while sampling points above or below the bounds are labeled with "1" and "-1", respectively.

LTPs could then be encoded in a histogram of 3^P different bins. However, this would result in extremely high dimensional feature vectors. For instance, a histogram of 6561 bins would be obtained by considering only $P = 8$ neighbors. Using $P = 16$ neighbors would result in 43046721 different labels. Thus, in practice LTPs are usually split into two different binary patterns, a "positive" and a "negative" part, as shown in Figure 2.6.

Regular histograms can then be obtained from both channels of the LTP descriptor separately. Both histograms are then concatenated to form the final texture representation.

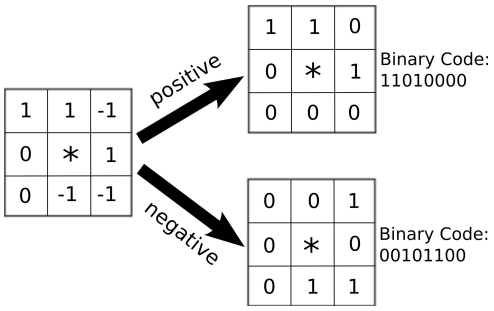


Figure 2.6.: ***Split of the LTP operator into a positive and negative part.*** Since using the ternary code directly would result in too high dimensional feature vectors, the basic LTP operator is split into a "positive" and a "negative" part. Thus, two regular LBP operators are obtained and histograms can be calculated from each LBP pattern separately which are subsequently concatenated to form the final descriptor.

A more detailed overview and explanation of LBP, LTP, as well as various extensions of both descriptors can be found in [49].

2.2.2. Local Descriptors

As stated earlier, opposed to global descriptors which are representations of the whole image, local descriptors only encode information about a restricted area around certain *keypoints* (also referred to as *interest points*). Thus, the detection of these keypoints in different scales is usually the first step of local feature extraction. For the application presented in this thesis, however, the keypoints can be calculated based on the positions of both eyes and the mouth (see section 4.2.1). Furthermore, in order to achieve scale invariance of the descriptor, keypoints are usually detected in the so-called scale-space, a multi-scale representation of an image. For the task at hand, this step can also be omitted since the size of the region around each interest point can be calculated based on the size of the cropped facial image after alignment (see Section 4.2.3.2 for details).

Hence, only the different feature extraction techniques are reviewed in more detail within this section. For the detection part of the local descriptors the reader is referred to the according publications [51, 52]. More detailed insights and comparison of different local keypoint descriptors can be found in [36].

Scale Invariant Feature Transform (SIFT)

In 2004, David Lowe presented a scale and rotation invariant local feature detector and descriptor called SIFT which gained much attention within the research community [51]. Nowadays, it still serves as a baseline for more recently developed keypoint descriptors.

A consistent orientation is assigned to every keypoint after interest point detection in the so-called Difference of Gaussian (DoG) scale-space in order to achieve rotation invariance of the descriptor. The gradient magnitudes and orientations are calculated for every pixel of a region around the detected keypoint in its according scale. An orientation histogram consisting of 36 bins covering 360° in 10° steps can then be formed in which maxima correspond to the dominant directions of the keypoint. Lowe found that assigning multiple orientations to one single interest point contribute significantly to the robustness of SIFT. Thus, multiple keypoints can be assigned to the same location and scale but with different orientations. More details about the orientation assignment can be found in [51]. After the dominant orientations were assigned, the gradient directions can be rotated relative to the main orientation of the keypoint.

For extraction of the final descriptor, first a 16×16 grid is created around each keypoint and the gradient magnitudes and directions of each image sample point are calculated. A 2D Gaussian window is additionally applied to put less emphasis on gradients far away from the actual interest point location. The gradient magnitudes and orientations of 4×4 subregions are then accumulated into 8 bin orientation histograms. The final SIFT descriptor is then given by concatenating the obtained histograms. Thus, the feature vector for every interest point consists of $4 \cdot 4 \cdot 8 = 128$ elements. The descriptor vector is then normalized to unit length to reduce the effects of illumination changes. For details about keypoint detection, orientation assignment and SIFT descriptor extraction the reader is referred to [51]. Figure 2.7 illustrates the process of SIFT descriptor generation.

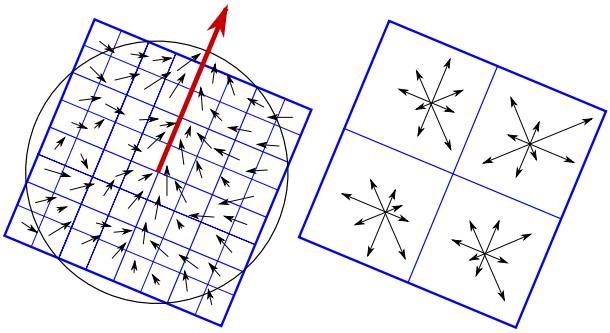


Figure 2.7.: ***SIFT feature extraction.*** The figure illustrates how SIFT descriptors are extracted from a previously detected keypoint. After the main direction has been assigned, all gradient orientations are rotated relative to the main orientation of the keypoint to achieve rotation invariance. Then, a rectangular region is created around each keypoint and gradient magnitudes and orientations are calculated for every image sample point. A 2D Gaussian window (black circle on the left) is applied to emphasize gradients near the center of the region which ensures smooth weighting. The gradients of 4×4 subregions are then accumulated into 8 bin orientation histograms, indicated on the right. Note that for illustration purposes a 2×2 descriptor array obtained from a region grid of size 8×8 is shown on the right hand side of the figure. For the actual SIFT descriptor 16×16 grid is constructed around each keypoint which is accumulated into 4×4 histograms. Image adapted from [51]

Speeded-Up Robust Features (SURF)

SURF is a fast and robust scale- and rotation-invariant interest point detector and descriptor, published by Bay *et al.* in 2008 [52] as an extension of SIFT.

As claimed by the authors, the standard version of SURF is several times faster, more compact and at the same time more robust against certain image transformations than comparable local descriptors such as SIFT. Similar to SIFT and its variants, SURF describes the distribution of intensity variation within a certain neighborhood around an interest point. However, instead of extracting the gradient information directly, SURF uses first order Haar wavelet responses in x and y-direction to approximate the gradient. For efficiency, SURF further exploits the concept of integral images, a quick and effective way of calculating the sum of pixel values in a given image or a rectangular subset of an image. This procedure drastically reduces processing time while at the same time improve the robustness of the resulting descriptor [52]. As done for SIFT, the first step of feature extraction is to identify a reproducible orientation for the interest point in order to increase the robustness against rotation. The dominant orientation can be found by calculating the sum of the Gaussian weighted Haar wavelet responses using a sliding window around a circular region around the interest point. The next step is to construct a square region of a fixed size symmetrically around the interest point. This region is then split into 4×4 sub-regions to preserve spatial information. Horizontal and vertical Haar wavelet responses, d_x and d_y, on 2×2 sub-divisions are subsequently calculated for each sub-region. Additionally, the obtained Haar wavelet responses are weighted with a Gaussian kernel which is centered at the particular interest point to compensate for geometric deformations. Thus, gradients closer to the center of the square region are treated more important than wavelet responses farther away from the interest point. The magnitudes as well as the directions of the weighted Haar wavelet responses are calculated, summed and concatenated to form the final feature vector. Hence, the descriptor for each sub-region is a four dimensional vector:

$$\boldsymbol{f}_{\text{SURF}} = \left[\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right] \tag{2.8}$$

Concatenation of $\boldsymbol{f}_{\text{SURF}}$ for all 4×4 sub-regions results in a descriptor of size 64 for every keypoint which is invariant to uniform illumination changes. Additional robustness to contrast changes is achieved by normalizing the final feature vector to have unit norm.

The process of local feature extraction is illustrated in Figure 2.8.

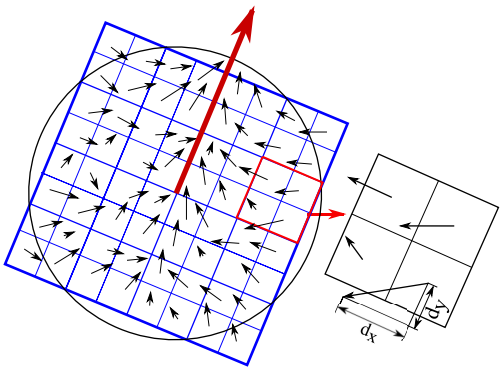


Figure 2.8.: ***SURF feature extraction.*** The figure illustrates the process of local feature extraction for SURF. First, the main orientation of a keypoint is found (read arrow) and the 4×4 sub-regions are rotated accordingly. Each sub-region is then split into 2×2 cells and the Haar-wavelet responses in x- and y-direction as well as their orientations are calculated in each cell. To compensate geometric deformations, the wavelet responses are additionally weighted with a Gaussian kernel centered at the interest point (black circle). Image adapted from [52]

2.3. Feature Space Transformations

In pattern recognition and face recognition in particular often feature space transformation techniques are applied because descriptors in its original size are usually too high dimensional to perform accurate and efficient classification in practice. Fortunately, it can be shown that the "intrinsic dimensionality" of the data is much lower. Thus, the high-dimensional feature vectors of size n can be projected into a lower dimensional subspace of size m ($m \ll n$) without noticeable loss of information.

The goal of feature space transformation techniques is to project the N high dimensional vectorized feature vectors $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ of size n into a lower dimensional subspace of size m using a unitary projection matrix $\boldsymbol{W} \in \mathbb{R}^{n \times m}$.

$$\boldsymbol{y}_k = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{x}_k; \quad \text{with} \quad \boldsymbol{x}_k \in \mathbb{R}^{n \times 1}, \quad \boldsymbol{y}_k \in \mathbb{R}^{m \times 1}, \quad m \ll n \tag{2.9}$$

The resulting feature vectors $\boldsymbol{y}_k \in \mathbb{R}^{m \times 1}$, with $k = 1, \cdots, N$, can then be used for classification. Three popular linear feature space transformation techniques are reviewed in this section: *Principal Component Analysis (PCA)*, *Linear Discriminant Analysis (LDA)*, and *Locality Preserving Projections (LPP)*. Note that also a number of non-linear techniques have been proposed to generate a low-dimensional representation from a high-dimensional space [53, 54, 55, 56, 57]. However, as stated in [58], although those methods achieve impressive results on artificial benchmark datasets, the obtained mappings are usually defined only on the training set and cannot easily be applied to unseen test data.

2.3.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [59, 60] aims to find a subspace whose basis vectors correspond to the maximum-variance directions of the original feature space. Thus, PCA tries to maximize the following objective function

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{PCA}} \boldsymbol{w}), \tag{2.10}$$

where the covariance matrix $\boldsymbol{S}_{\mathrm{PCA}}$ is defined as

$$\boldsymbol{S}_{\mathrm{PCA}} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}. \tag{2.11}$$

The output set of principal vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m]$ is an orthonormal set of vectors representing the eigenvectors of the sample covariance matrix associated with the $m \ll n$ largest eigenvalues. All columns of $\boldsymbol{X}$ should be normalized to unit norm before extracting the eigenvectors of $\boldsymbol{S}_{\mathrm{PCA}}$. Furthermore, it has to be ensured that $\boldsymbol{X}$ is zero mean, i.e. the mean of all samples has to be subtracted from all normalized feature vectors.

A classic application of PCA is the *Eigenfaces* method for face recognition presented by Turk and Pentland in [61] (see Section 3.3 for details). However, in image processing the covariance matrix $\boldsymbol{S}_{\mathrm{PCA}}$ is often too large for efficient and fast eigenvalue decomposition since $n \gg N$. To overcome this difficulty it was shown in [61] that $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ can be replaced with its surrogate $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ whose eigenvectors can be calculated much more efficiently due to its lower dimensionality.

2.3.2. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [59, 60] aims to preserve the discriminating information between the classes. Again, a set of n-dimensional samples $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ each associated with one of $i = 1, \cdots, C$ classes is considered, where N_i denotes the number of samples of class c_i. The between-class scatter matrix $\boldsymbol{S}_{\mathrm{b}}$ and the within-class scatter matrix $\boldsymbol{S}_{\mathrm{w}}$ are defined as

$$\boldsymbol{S}_{\mathrm{b}} = \frac{1}{N} \sum_{i=1}^{C} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}} \tag{2.12}$$

and

$$\boldsymbol{S}_{\mathrm{w}} = \frac{1}{N} \sum_{i=1}^{C} \left[\sum_{j=1}^{N_i} (\boldsymbol{x}_j^{(i)} - \boldsymbol{\mu}_i)(\boldsymbol{x}_j^{(i)} - \boldsymbol{\mu}_i)^{\mathrm{T}} \right], \tag{2.13}$$

respectively, where $\boldsymbol{\mu}_i$ is the mean of all images of class c_i, $\boldsymbol{\mu}$ is the total sample mean vector and $\boldsymbol{x}_j^{(i)}$ is the j-th sample of class i.

LDA solves the Fisher criterion, i.e. the projection is chosen such that the variance between the classes is maximized while the variance within each class is minimized. Consequently, the objective function is as follows:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \left(\frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w}} \right). \tag{2.14}$$

The optimal projection basis for LDA is the set of generalized eigenvectors associated with the m largest eigenvalues of $\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$. Note that there are at most $C-1$ nonzero generalized eigenvectors. Hence, an upper bound for the dimensionality of the transformed subspace exists, where $m \leq C-1$. Furthermore, at least $n+C$ samples are required to ensure that $\boldsymbol{S}_{\mathrm{w}}$ does not become singular. In practice, however, this often cannot be achieved. To overcome the problem of the singularity of $\boldsymbol{S}_{\mathrm{w}}$, Belhumeur *et al.* [62] proposed to first project the features into a lower dimensional subspace by applying PCA.

One of the best known applications of LDA is the *Fisherfaces* method developed by Belhumeur *et al.* in [62] (see Section 3.3 for details).

Due to the fact that LDA uses the class information directly, it outperforms PCA in many applictions. On the other hand, Martinez *et al.* empirically showed in [60] that PCA can outperform LDA when the training set is small and therefore cannot represent the underlying class distribution.

2.3.3. Locality Preserving Projections (LPP)

One major disadvantage of PCA and LDA is that both methods only see the the global euclidean structure of the feature space. It is known, however, that for many pattern recognition problems, especially for face recognition, the intrinsic structure of the feature space is actually non-linear [63, 64, 56, 65, 66]. To overcome these limitations, generalized versions of PCA and LDA were proposed which apply a so called kernel trick for nonlinear embedding [67, 68]. However, these methods do not explicitly consider the special manifold structure of the feature space.

Locality Preserving Projections (LPP), proposed by He *et al.* in [58], assumes that the feature vectors reside on a nonlinear submanifold hidden in the original feature space. Unlike PCA or LDA, LPP tries to find an embedding that preserves local information. Thus, samples that are close in the original feature space should be close in the projected subspace. On the other hand, if samples are located far away in the original feature space, they should be mapped far apart in the transformed space. Thus, LPP aims at obtaining a subspace that best preserves the local manifold structure of the original feature space. This is achieved by modeling the manifold structure of the feature space using a nearest-neighbor graph.

First, an adjacency graph $\boldsymbol{\mathcal{G}}$ is defined and an edge is put between two nodes k and j if they are "close", i.e. if they are within an ϵ-neighborhood[1]. LPP then tries to optimally preserve this graph when choosing projections. After constructing the graph, weights have to be assigned to the edges. Therefore, a sparse symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ is created with $\boldsymbol{S}_{k,j}$ having the weight of the edge joining vertices k and j, and 0 otherwise. These weights are often calculated by using a heat-kernel for instance:

$$\boldsymbol{S}_{k,j} = \begin{cases} e^{-\frac{\|\boldsymbol{x}_k - \boldsymbol{x}_j\|^2}{2\sigma^2}}, & \text{if } \|\boldsymbol{x}_k - \boldsymbol{x}_j\|^2 < \epsilon \\ 0, & \text{otherwise.} \end{cases} \tag{2.15}$$

Here, σ denotes a constant factor for normalization and ϵ is a user defined threshold which defines the maximal distance between two samples. LPP tries to minimize the objective function

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} \sum_{kj} (\boldsymbol{y}_k - \boldsymbol{y}_j)^2 \boldsymbol{S}_{k,j} \\ &= \arg\min_{\boldsymbol{w}} \sum_{kj} (\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_k - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_j)^2 \boldsymbol{S}_{k,j}. \end{aligned} \tag{2.16}$$

Following some simple algebraic steps, it is possible to show that equation 2.16 finally results in a generalized eigenvalue problem:

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w} = \lambda \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}, \tag{2.17}$$

where $\boldsymbol{D}$ is a diagonal matrix whose entries are column (or row since $\boldsymbol{S}$ is symmetric) sums of $\boldsymbol{S}$, i.e. $\boldsymbol{D}_{kk} = \sum_j \boldsymbol{S}_{kj}$. Furthermore, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$ is the so called Laplacian matrix [53]. The mathematical proof of equation 2.17 as well as the theoretical justifications of LPP can be found in [58]. The projection matrix $\boldsymbol{W}$ is constructed by concatenating the solution to the above equation, i.e. the column vectors of $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m]$ are ordered ascendingly according to their eigenvalues.

It has been shown in [69] that LPP can be applied to the field of appearance based face recognition and outperforms *Eigenfaces* and *Fisherfaces* on various benchmark datasets. This method is referred to as *Laplacianfaces* (see Section 3.3) for the remainder of this thesis.

[1] Instead of connecting two nodes based on their ϵ-neighborhood, another possibility suggested by [58] is to utilize the k-Nearest-Neighbor (k-NN) scheme. Here, two nodes are connected by an edge if they are among the k nearest neighbors of each other.

2.4. Classification

2.4.1. Classification using k-Nearest-Neighbor (k-NN)

One of the simplest yet widely used techniques for classification is the k-Nearest-Neighbor (k-NN) method, which is purely memory based and requires no model to be fit during training [70]. First, the distances between an unknown test sample $\boldsymbol{t}$ and all training samples in a common feature space are calculated. In its simplest form, the nearest neighbor (NN) classification, the test sample is then assigned to the class that belongs to the training sample that has the smallest distance to $\boldsymbol{t}$. Figure 2.9 illustrates the principle of nearest neighbor classification. In this example the test sample is classified as class 1 when NN is applied.

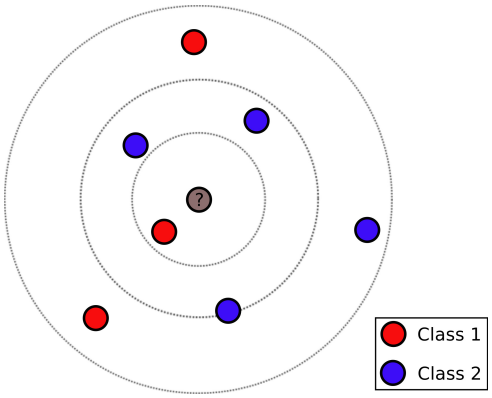


Figure 2.9.: ***Illustration of the k-NN classification scheme.*** A test sample $\boldsymbol{t}$ (marked with "?") is classified as class 1 if a nearest-neighbor classifier is applied. For k-NN with $k = 5$ the test instance is assigned to class 2.

Several different distance measures were proposed in the past. Although the Euclidean distance d_E between two m-dimensional feature vectors has traditionally been used for classification

$$d_E(\boldsymbol{t}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} (\boldsymbol{t}(i) - \boldsymbol{y}(i))^2} = \|\boldsymbol{t} - \boldsymbol{y}\|_2, \tag{2.18}$$

a number of other distance metrics are often applied in face recognition and other pattern recognition problems including the Manhatten-distance d_1

$$d_1(\boldsymbol{t}, \boldsymbol{y}) = \sum_{i=1}^{m} |\boldsymbol{t}(i) - \boldsymbol{y}(i)| = \|\boldsymbol{t} - \boldsymbol{y}\|_1, \tag{2.19}$$

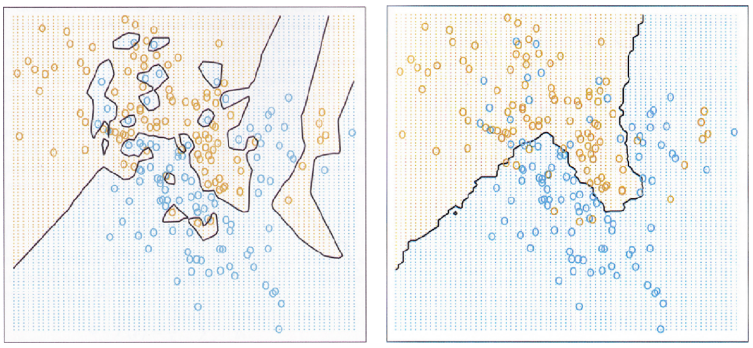
or the χ^2-distance d_χ

$$d_{\chi^2}(\boldsymbol{t}, \boldsymbol{y}) = \frac{1}{2} \sum_{i=1}^{m} \frac{(\boldsymbol{t}(i) - \boldsymbol{y}(i))^2}{(\boldsymbol{t}(i) + \boldsymbol{y}(i))}, \tag{2.20}$$

which is particularly useful when comparing histograms. In case multiple training instances of different classes have the same distance to the test sample, the class of $\boldsymbol{t}$ is assigned randomly to any of those classes.

A widely used extension of the nearest neighbor classification scheme is the k-NN method. Again, the first step is to calculate the distances between the test instance $\boldsymbol{t}$ and all training samples. All distances are then sorted ascendingly and the class assignment of $\boldsymbol{t}$ is done by using a majority voting between the k closest training points, where k is a user-defined parameter. In the example from Figure 2.9, three objects of class 2 and two objects of class 1 are found in the neighborhood of $\boldsymbol{t}$ using a 5-NN classification paradigm. Therefore, the test sample is classified as class 2.

The choice of k is critical to achieve good classification performances and highly depends on the statistical distribution of the data. Large values of k tend to decrease the classifier's sensitivity to training samples falling in the area of an adjacent class. However, simultaneously the selectivity between classes might be reduced [70]. Figure 2.10 illustrates this behavior of k-NN for different values of k. The decision boundary for $k = 15$ is fairly smooth compared to a classification where $k = 1$ was used.



(a) $k = 1$ (b) $k = 15$

Figure 2.10.: ***Comparison of the k-NN classifier for different values of k.*** The figure compares the decision boundaries of the k-NN classifier for $k = 1$ (a) and $k = 15$ (b) on the same data. The decision boundary for $k = 15$ appears to be smoother than for $k = 1$ which shows that larger values for k decrease the sensitivity of the classifier to training samples falling in the area of an adjacent class. However, using larger values for k might also reduce the selectivity between classes. Image source: [70]

Despite its simplicity, k-NN has been successfully applied in a variety of classification problems including face recognition (see Section 3.3).

2.4.2. Sparse Representation Classification (SRC)

Sparse Representation Classification (SRC) is based on Compressed Sensing (CS), a mathematical framework for signal measurement and reconstruction. Recently, SRC has been successfully applied to face recognition and promising results were obtained even under difficult lighting conditions and partial occlusion [71, 48] (see Section 3.3). SRC assumes that given a sufficiently large number of training samples a test vector can be represented as a linear combination of training instances of the same class.

Let $\boldsymbol{A} \in \mathbb{R}^{m \times l}$ be the normalized matrix of training samples transformed into the m-dimensional feature space and $\boldsymbol{t} \in \mathbb{R}^m$ be the normalized transformed feature vector of the test image, where m is the dimensionality of the feature space and l the number of training samples. The matrix of training samples of the i-th object class is given by

$$\boldsymbol{A}_i = [\boldsymbol{a}_{i,1}, \boldsymbol{a}_{i,2}, \cdots, \boldsymbol{a}_{i,l_i}] \in \mathbb{R}^{m \times l_i}, \tag{2.21}$$

where l_i is the number of training samples of class i. Thus, $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_C] = [\boldsymbol{a}_{1,1}, \boldsymbol{a}_{1,2}, \cdots, \boldsymbol{a}_{1,l_1}, \cdots, \boldsymbol{a}_{C,1}, \cdots, \boldsymbol{a}_{C,l_C}]. \tag{2.22}$$

SRC assumes that any test instance $\boldsymbol{t}$ reside in the linear span of the training samples of the i-th object class

$$\boldsymbol{t} = \alpha_{i,1}\boldsymbol{a}_{i,1} + \alpha_{i,2}\boldsymbol{a}_{i,2} + \cdots + \alpha_{i,l_i}\boldsymbol{a}_{i,l_i} \tag{2.23}$$

for some scalar weights $\alpha_{i,j} \in \mathbb{R}$ with $j = 1, \cdots, l_i$. In other words, SRC tries to represent the feature vector of the test image $\boldsymbol{t}$ as a linear combination of the training samples of the same class. This can be expressed as

$$\boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}_0, \tag{2.24}$$

where $\boldsymbol{p}_0 \in \mathbb{R}^l$ is a sparse coefficient vector whose entries only associated with the i-th class should be 1 and the rest 0

$$\boldsymbol{p}_0 = [0, 0, \cdots, \alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,l_i}, \cdots, 0, 0]^{\mathrm{T}}. \tag{2.25}$$

Therefore, the class of a test sample can simply be found by solving a system of linear equations. However, typically the number of features after transformation is much lower than the number of observations, i.e. $m < l$. Hence, it is obvious that equation 2.24 is underdetermined which means that its solution is not unique.

A common way to solve this problem is to find the solution that minimizes the ℓ^2-norm:

$$(\ell^2) \quad \hat{\boldsymbol{p}}_2 = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\|_2 \quad \text{subject to} \quad \boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}. \tag{2.26}$$

However, it was shown in [71] that for this solution $\hat{\boldsymbol{p}}_2$ is generally dense which results in large non-zero entries associated with different classes. Due to the observation that each test sample can be expressed as a linear combination of training instances of the same class, the objective of SRC is to find the sparsest solution of equation 2.24

$$(\ell^0) \quad \hat{\boldsymbol{p}}_0 = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\|_0 \quad \text{subject to} \quad \boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}. \tag{2.27}$$

Note that the ℓ^0-norm $\|\cdot\|_0$ is a pseudonorm and simply counts the number of non-zero entries in a vector. Unfortunately, finding the sparsest solution of an underdetermined system of linear equations is known to be NP-hard, i.e. it cannot be solved in polynomial time [72]. However, Compressed Sensing (CS) theory suggests that the ℓ_0-norm solution of equation 2.27 can be replaced by solving a convex optimization problem via ℓ^1-norm minimization

$$(\ell^1) \quad \hat{\boldsymbol{p}}_1 = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\|_1 \quad \text{subject to} \quad \boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}. \tag{2.28}$$

Ideally, the nonzero entries in the sparse coefficient vector $\hat{\boldsymbol{p}}_1$ will all be associated with the columns of $\boldsymbol{A}$ which represent a single class. However, in real-world applications noise and modeling error may lead to small nonzero entries associated with different object classes.

Thus, the minimal residual $r_i(\boldsymbol{t})$ between $\boldsymbol{t}$ and $\boldsymbol{A}(\boldsymbol{\delta}_i \odot \hat{\boldsymbol{p}}_1)$ is chosen to indicate the class the test vector $\boldsymbol{t}$ belongs to:

$$\mathrm{ID}(\boldsymbol{t}) = \arg\min_i r_i(\boldsymbol{t}) \quad \text{with} \quad r_i(\boldsymbol{t}) = \|\boldsymbol{t} - \boldsymbol{A}(\boldsymbol{\delta}_i \odot \hat{\boldsymbol{p}}_1)\|_2, \tag{2.29}$$

where $\odot$ denotes the elementwise multiplication known as Hadamardt-Schur product. The vector $\boldsymbol{\delta}_i \in \mathbb{R}^l$ is called the characteristic function of class i. It can be thought of as a filter vector which is 1 for all training samples of class i and 0 elsewhere. A detailed description of CS theory and SRC can be found in [71, 73, 48].

2.4.3. Support Vector Machines (SVMs)

Introduced by Vapnik *et al.* [74] in 1995, Support Vector Machines(SVMs) became one of the best known and well established tools for classification and supervised machine learning over the past two decades. A brief overview of the basic ideas of SVMs is given in this section. For a more detailed description and mathematical justification of SVMs the interested reader is referred to the books of Vapnik *et al.* [74, 75]. Furthermore, a significant amount of literature which gives a detailed introduction of supervised machine learning techniques and SVMs in particular was published during the past years [76, 77, 78].

SVM in its original implementation was introduced as binary classifier. Given a set of training examples $\boldsymbol{y}_i \in \mathbb{R}^m$ $(i = 1, \cdots, l)$ each belonging to one of two classes $\boldsymbol{c}(\boldsymbol{y}_i) \in \{1, -1\}$, the objective of SVM is to find a hyperplane which separates the positive from the negative examples. Since many such hyperplanes exists, the most reasonable choice is to find the one with the *largest margin* between the two classes. Let d_+ be the shortest distance between the separating hyperplane and the closest positive training instance. The shortest distance between the separating hyperplane and the closest negative sample is denoted by d_-. Then the margin is defined as $d_+ + d_-$ [76]. The training samples that lie on the margin are called the *support vectors*. A hyperplane is characterized by its normal vector $\boldsymbol{n}$ and the perpendicular distance of the plane to the origin $b/\|\boldsymbol{n}\|$, where $\|\boldsymbol{n}\|$ is the Euclidean norm of $\boldsymbol{n}$. If the data is separable, the hyperplane with the largest margin between the two classes can be found by solving the following minimization problem:

$$\arg\min_{\boldsymbol{n},b} \frac{1}{2}\|\boldsymbol{n}\|^2 \quad \text{subject to} \quad \boldsymbol{c}(\boldsymbol{y}_i)(\boldsymbol{n} \cdot \boldsymbol{y}_i + b \geq 1). \tag{2.30}$$

Once an SVM was trained, i.e. a hyperplane was found by solving the minimization problem in equation 2.30, it has to be decided on which side of the decision boundary a given test sample $\boldsymbol{t}$ is located. Thus, the class label of the test sample is given by $\operatorname{sgn}(\boldsymbol{n} \cdot \boldsymbol{t} + b)$, where $\operatorname{sgn}(\cdot)$ is the signum function.

Applying the minimization problem described in equation 2.30 to non-separable data will not lead to a feasible solution. To deal with non-separable data, Cortes and Vapnik [75] propose to relax the condition of equation 2.30 by introducing positive slack variables ξ_i, which measure the degree of misclassification of the data $\boldsymbol{y}_i$. A natural way to assign extra costs for errors is to extend the optimization problem in equation 2.30 to

$$\arg\min_{\boldsymbol{n},b,\xi} \left\{ \frac{1}{2}\|\boldsymbol{n}\|^2 + C\sum_{i=1}^{l} \xi_i \right\} \quad \text{subject to} \quad \boldsymbol{c}(\boldsymbol{y}_i)(\boldsymbol{n} \cdot \boldsymbol{y}_i + b \geq 1 - \xi_i), \quad \xi_i \geq 0, \tag{2.31}$$

where C is a parameter defined by the user. A large C corresponds to a high penalty of errors. Figure 2.11 compares the linear separating hyperplanes of an SVM for the separable (a) and the non-separable case (b).

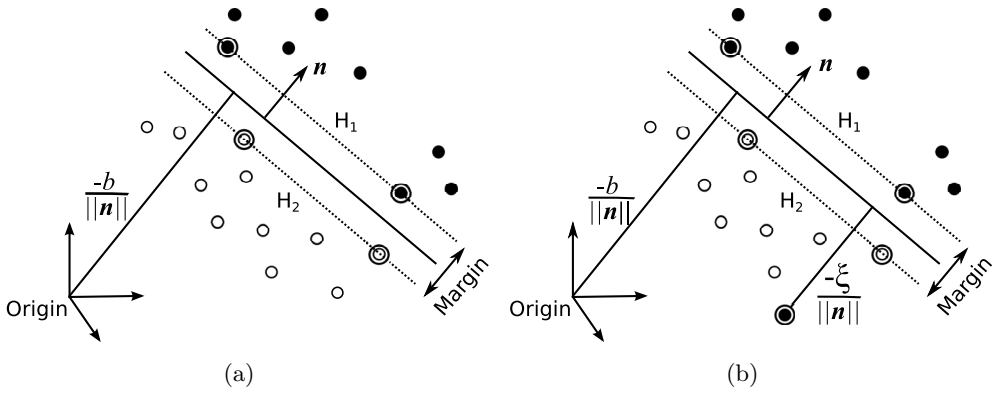


Figure 2.11.: ***Comparison of separable vs. non-separable data.*** This figure illustrates separating hyperplanes of an SVM for separable data (a) and non-separable data (b). Positive slack-variables ξ_i are introduced for the non-separable case which measure the degree of misclassification of the data. Furthermore, a user-defined penalty parameter C is defined which controls how much misclassifications are penalized. A large C corresponds to a high penalty of errors.

In practice, data of two different classes often cannot be separated adequately by a linear function. However, [79] showed that a so called kernel trick, originally proposed by Aizerman *et al.* [80] in 1964, can be used to transform the data into a higher dimensional space. A mapping function $\Phi : \mathbb{R}^m \mapsto \boldsymbol{\mathcal{H}}$ can be applied in order to subsequently fit a maximum-margin hyperplane in the higher dimensional feature space $\boldsymbol{\mathcal{H}}$. A variety of different kernel functions were proposed by researchers such as linear kernels, polynomial kernels, and the Radial Basis Function (RBF) kernel [76, 78]. The interested reader is referred to [76] for a detailed explanation of non-linear SVMs.

Figure 2.12 illustrates the process of transforming the non-linearly separable data into a higher dimensional feature space where the two classes can be separated by a linear hyperplane.

As mentioned earlier, SVM in its original implementation is a binary classifier which can only handle two-class problems. Several extensions were proposed in order to handle multi-class problems. The most common way is to split the multi-class problem in multiple binary classification problems [82, 83], where either a *one-versus-all* strategy or *one-versus-one* technique is applied to get a final result. On the other hand, Crammer and Singer [84] proposed to treat multi-class classification as one single optimization problem rather than decomposing it into multiple binary classifications. For details and comparisons between multi-class SVM strategies the reader is referred to [82, 83].

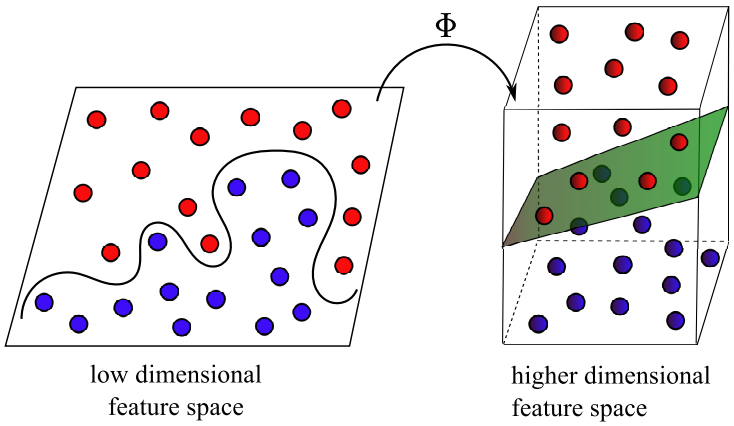


Figure 2.12.: ***Kernel trick for non-linear SVMs.*** The samples of the two classes (blue and red circles) cannot be separated by a linear hyperplane in the low dimensional feature space. Therefore, a kernel function Φ was proposed by [79] to efficiently transform the data into a higher dimensional space where the two classes are linearly separable. Image inspired by [81]

2.5. Chapter Summary

In this chapter, state-of-the-art algorithms commonly used in pattern recognition were introduced and reviewed. In particular, techniques considered to be relevant for understanding the proposed PRF as well as the face recognition systems used for benchmarking were emphasized. Many techniques for non-invasive individual identification, including the presented PRF, are based a standard pattern recognition pipeline: *Feature Extraction*, *Feature Space Transformation*, and *Classification*.

Hence, in Section 2.2 global feature extraction techniques as well as local key point descriptors were introduced which have been proven to be efficient visual representations of image and video content. In particular, global descriptors such as Gabor-based features and LBP as well its extensions are known to perform well for human face recognition and were therefore reviewed in more detail. To cope with disadvantages originating from the holistic nature of global features, a large and growing amount of literature proposed and investigated local keypoint descriptors. Two sophisticated local feature extraction techniques, SIFT and SURF, were thus briefly discussed in Section 2.2 as well.

Many feature extraction techniques result in extremely high-dimensional descriptors. Thus, feature space transformation techniques are required for many applications in order to perform robust and fast recognition in practice. The goal of feature space transformation is to reduce the dimensionality of the feature space without losing information that is important for discrimination. Three of the most commonly applied algorithms in practical applications were reviewed and compared in Section 2.3: PCA, LDA, and LPP.

Finally, in Section 2.4, three of the most widely applied supervised classification techniques were discussed. While k-NN and SVM are traditional methods to assign one particular class to a test sample, more recently a novel classification paradigm known as SRC was proposed. SRC is based on Compressed Sensing (CS) theory, a mathematical framework for signal measurement and reconstruction. SRC is utilized within the proposed face recognition framework for great apes and has been proven to perform well for other pattern recognition task as well.

3. State of the Art

3.1. Chapter Overview

Recently, the development of automatic routine procedures to efficiently process huge amounts of biologically relevant audiovisual footage attracts the attention of many researchers, engineers as well as computer scientists. The emerging and highly interdisciplinary field of *Animal Biometrics* aims at developing and applying approaches to automatically detect, represent, and interpret phenotypic appearance of various animal species [85]. Such algorithms can be used to detect animals in audiovisual footage, recognize particular behaviors, classify different species, or even identify individuals. Many biologically motivated research areas, such as ecology, behavioral research, or estimation of population sizes would benefit from automatic methods for non-invasive animal monitoring. Due to its complexity, the field of *Animal Biometrics* requires close interaction of many scientific disciplines. While biologists have detailed knowledge about the visual cues that let humans differentiate between different animal species, sub-species or individuals, computer scientists need this biological expertise in order to implement applicable computer algorithms. This chapter gives an extensive and representative literature review of state-of-the-art approaches for non-invasive biometric animal monitoring. This includes algorithms for detection, tracking and behavior analysis of animals as well as methods for individual identification and species recognition. Although the field of *Animal Biometrics* includes both, video and audio based applications (e.g. [86, 87, 88, 89]), the following chapter focuses on image and video based systems in particular.

Starting from the assumption that humans and our closest relatives, the great apes, share similar properties of the face the Primate Recognition Framework (PRF) proposed in this thesis is based on face detection and recognition algorithms to differentiate between different individuals. Hence, the second part of this chapter gives an overview of state-of-the-art face recognition algorithms originally developed to identify humans in images or videos.

3.2. Visual Animal Biometrics

While the development of automatic approaches to detect, track, identify, and analyze the behavior of human beings has been an active research topic for decades, *Animal Biometrics* is a rather new field of research. Early work concentrated on intrusive methods in which animals are physically marked using ear tags, passive integrated transponder (PIT) tags, leg bands or dyes [27, 28]. In captive environments often radio frequency identification (RFID) chips or spe-

cial hardware such as sensors for active infrared, laser or microwave radio signals have been proven to be useful for animal monitoring [90]. However, wildlife environments often impose immense demands on hardware and software due to limited power supply and harsh environmental conditions such as heating by strong sunlight, cold and icing in high mountainous locations or corrosion in moist regions. Therefore, such invasive hardware-based animal monitoring techniques are typically used in captive environments for livestock breeding. Moreover, Schatzmann argues in [91] that invasive methods such as branding and transponder implantation can cause permanent pain and severe complications. Another study by Edwards *et al.* [92] showed that ear tagging often results in inflammatory response causing extreme discomfort and pain. Furthermore, especially in wildlife settings, invasive methods can be a severe intrusion into ecosystems itself and may therefore bias observation due to potential changes in behavior, reproduction, and even survival [29, 30, 31]. In addition, some taxonomic groups might be difficult to tag due to small body size and constraints of field conditions or may not retain tags long enough to be useful for behavioral ecological research [93].

On the other hand, non-invasive methods for analyzing *Animal Biometrics* can be extremely ambitious due to different lighting conditions, various body postures, partial occlusion, camouflaged animals, and many more challenging conditions frequently present in real-world settings. However, the development of automatic routine procedures for non-invasive animal monitoring and analysis of remotely gathered audiovisual recordings is a prerequisite for many biological fields such as ecology, biodiversity conservation, and behavior understanding. Furthermore, non-invasive methods have several advantages over intrusive methods because animals are neither harmed nor affected in their behavior by applying autonomous video surveillance cameras.

3.2.1. Animal Detection and Tracking

During the past 20 years a large and growing amount of literature describing techniques for different content-based image and video retrieval tasks has been published. Especially the retrieval of animal pictures has recently attracted the attention of computer scientists [94, 95]. While Schmid *et al.* [94] solely exploit texture information in combination with unsupervised clustering techniques to retrieve images showing animals of the same kind, Berg and Forsyth [95] additionally incorporate other cues such as color, shape, and metadata from Google's text search to identify images containing categories of animals.

A different field of research is the automatic detection of animals in images or videos. Opposed to image retrieval, animal detection aims at determining the locations and sizes of animals within an image or video. Automatic approaches for animal detection build the basis of subsequent tasks such as tracking, behavior analysis, species recognition or individual identification.

Many approaches that either detect segments of the animals' body (e.g. the face) or complete bodies in image or video sequences can be found in the literature. The following section gives an overview of state-of-the-art systems for automatic animal detection, tracking, and behavior analysis.

Automatic animal detectors can be classified into five main categories: *Template Matching*, *Rigid Object Detectors*, *Local Keypoint Detectors*, *Model-Based Detectors*, and *Motion-Based Detectors*.

Template Matching

The simplest animal detectors use template matching to localize animals of a specific species in images or videos. One prototypical instance of the object's visual appearance is used to detected the animal of interest. A template image is typically generated by taking either only one or the mean of many example images that best represent the desired object. The template is then shifted across the source image. By calculating certain similarity metrics such as normalized cross-correlation or intensity difference measures [96] the object location is defined as the area corresponding to the highest similarity score. Kastberger *et al.* [97] for instance apply a template matching technique to detect and track individual agents in densely packed clusters of giant honey bees. The authors use a 3D stereoscopic imaging method to study behavioral aspects of shimmering, a defense strategy of bee collectives. After detection, stereo matching, and tracking of individual bees, 3D motion patterns were analyzed using luminance changes in subsequent frames.

Although template matching algorithms might be fast and easy to implement, they only achieve adequate results in rather controlled settings and often perform poorly in natural wildlife environments due to cluttered background and geometrical deformations of the object to be detected. To achieve invariance of the method against object deformations, different scales and rotations must be applied which significantly increase processing time. Hence, more sophisticated object localization algorithms must be applied to automatically detect free-living animals in their natural habitats.

Rigid Object Detectors

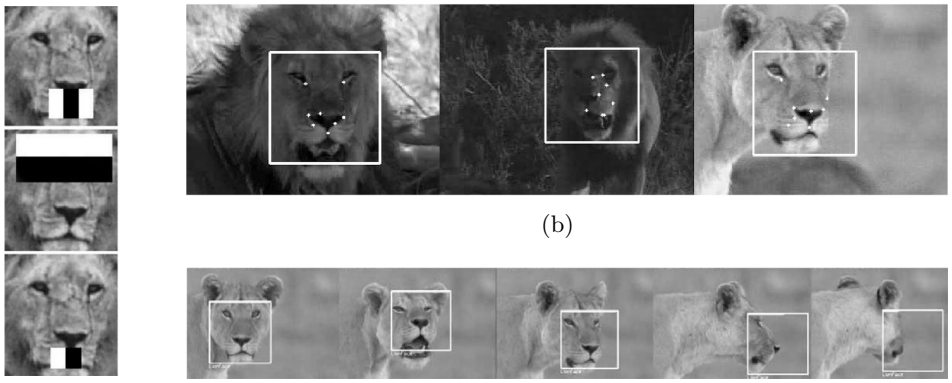
A more sophisticated approach to detect objects in natural scenes is to apply so-called rigid object detectors. As the name suggests, these kind of detectors are limited to non-deformable objects with similar shape. Visual descriptors and the spatial relationship between them are utilized to localize rigid objects, where features vary among instances but their spatial configuration remains similar. For human face and pedestrian detection, rigid object detectors have

been used for more than a decade. Although Rowley *et al.* [98] already achieved promising results with a neural-network based human face detection system in 1998, the probably best known algorithm for rigid object detection was presented by Viola and Jones in 2001 [99]. It uses a boosted cascade of simple Haar-like features which exploit local contrast configurations of an object in order to detect a face. The AdaBoost algorithm [100] is utilized for feature selection and learning. Numerous improvements have been proposed over the last decade to achieve wider pose invariance, most of them relying on the same principles as suggested by Viola and Jones. For an overview of face detectors the reader is referred to [101, 102, 103]. Later, Dalal and Triggs [104] proposed to use Histograms of Oriented Gradients (HOG) and a linear Support Vector Machine (SVM) for the detection of humans in low-resolution images. They showed that locally normalized HOG descriptors provide a robust feature set to reliably detect pedestrians, i.e. humans in upright positions, even in cluttered backgrounds, difficult lighting conditions, and various poses.

After the successful application of these techniques to the field of human detection in natural scenes, computer scientists started to adapt and extend these ideas to detect animals in images and videos as well.

Burghardt *et al.* for instance proposed a system for lion detection, tracking, and basic locomotive behavior analysis in [105, 106]. Although the authors use an enlarged set of Haar-like features (see Figure 3.1(a)), the initial face detection stage is based upon the original approach introduced by Viola and Jones [99]. Once a face has been localized in a frame, the Lucas-Kanade-Tomasi method [107] is applied to track the face region through the video sequence using a number of interest points on the lion's face (see Figure 3.1(b)). Furthermore, a rectangular interest model was created for locations where a face was spotted to achieve accurate and temporal coherent tracking performance. Figure 3.1(c) shows the detection and tracking results for an example image sequence. Information of the detection and tracking process was used to classify basic locomotive behavior of the animal. More explicitly, the vertical component of the head trajectory was used to detect actions such as walking, trotting, and standing.

The approach of Viola and Jones [99] has also actively been studied to detect heads of cat-like animals in images [108, 109, 110, 111]. However, in [108, 109] the authors argue that applying algorithms for human face detection directly to detect the head of cat-like animals would perform poorly. Since cat faces have a more complicated texture than humans faces and the shape of cat heads can vary significantly from individual to individual, the extraction of features would result in a high intra-class variance which is crucial for most detection algorithms. Zhang *et al.* overcome this burden in [108] by jointly utilizing texture and shape by applying different alignment strategies to separately train a shape and a texture detector. The shape information is kept by aligning the cat face such that the distance between the two ears is the same through



(a) (b) (c)

Figure 3.1.: ***Detection and tracking of lion faces.*** (a) Three most characteristic Haar-like features used to detect lion faces. (b) Chosen interest points used for tracking of lion faces through the video sequence. (c) Results of the proposed lion face detection and tracking algorithm through various poses. Image Source: [105, 106]

the entire training set. On the other hand, by aligning the faces according to their eyes the texture information is preserved while the shape of the cat's head is blurred. Figure 3.2(a) shows the mean image of the used training data for different alignment strategies. Furthermore, they use a set of Haar-like features on oriented gradients (see Figure 3.2(b)) as features which was shown to outperform other descriptors commonly used for face detection. In a second step, they jointly train a final classifier to fuse the outputs of the initial shape and texture detectors. In the detection phase, first both detectors are applied separately by using a sliding window to get initial face locations. A final decision is made by applying the joint shape and texture fusion classifier. The same authors apply their algorithm to other cat-like species like tigers, cheetahs, and pandas in [109]. To further handle the misalignment cost between both detectors they present a novel deformable detection approach which considers both misalignment cost and the outputs of the shape detector and texture detector. Also Kozakaya *et al.* use a two step approach to detect cat faces in images [110]. Opposed to the work done by Zhang *et al.* , they do not use two different alignment strategies but rather extract complementary feature sets to gather shape and texture information. In a first step a candidate search is performed which uses simple Haar-like features and AdaBoost as done by Viola and Jones in [99]. Although the approach by Viola and Jones is fast to compute and easy to implement, it often is not discriminative enough to deal with complicated shape and texture. Therefore, the authors suggest to use more sophisticated features in a second phase to verify face candidates. Thus, Kozakaya *et al.* use Co-occurrence Histograms of Oriented Gradients (CoHOG) [112] since they have strong classification capability to represent various cat face patterns. Due to the high dimensionality of CoHOG, a simple linear classifier obtained by a linear SVM is used for candidate verification.

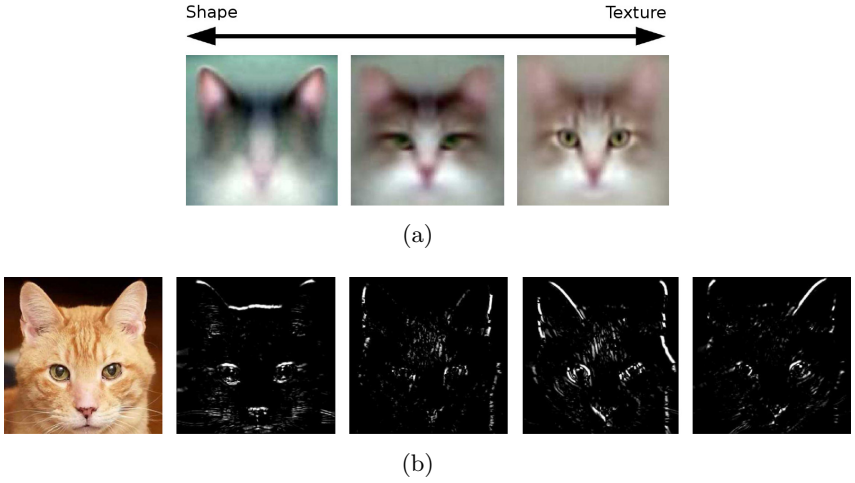


Figure 3.2.: ***Detection of cat faces in images.*** (a) Depending on the alignment, shape or texture information can be extracted. More shape information is kept by aligning the faces by ears (left), while more texture information is kept if the faces are aligned by their eyes (b) Orientation of the gradient information in four different directions used as feature set for the detection of cat faces. Image Source: [108, 109]

They evaluated their approach on the dataset provided by Zhang *et al.* in [108]. However, the experimental conditions were not strictly the same since the evaluation paradigm differs from the one used in [108]. Nevertheless, the results suggest that the proposed approach outperforms the method by Zhang *et al.* significantly. Another advantage of this method is that the approach is much more generic since it is not restricted to detect faces of cat-like animals only but other rigid objects as well.

Starting from the assumption that humans and our closest relatives share similar properties of the face, rigid object detectors have recently also been used to detect faces of African great apes [113]. The method by Ernst and Küblbeck is based on ideas of the approach by Viola and Jones [99] but was improved significantly by using multiple consecutive classification stages with increasing complexity and different illumination invariant feature sets to detect faces of chimpanzees and gorillas in images and video sequences. Each stage is built out of a feature extraction step and a classifier. One out of three illumination invariant features can be applied: edge orientation features, census features [114], and structure features built out of scaled versions of census features. Real-time capability is achieved by using simple and fast pixel-based features in the first stages and more sophisticated and therefore more complex descriptors in subsequent stages. Each stage consists of look-up tables that were build in an offline training procedure using Real-AdaBoost [115]. The first stages can be considered as a fast but inaccurate candidate search while the remaining stages focus on slower but more accurate classification. Although the proposed system has some robustness to difficult lighting situations, it lacks in invariance to

severe occlusion and far-off frontal poses. However, for subsequent analysis such as individual identification for instance, faces often are expected to be in a full-frontal pose. Therefore, the approach by Ernst and Küblbeck [113] has been used in this thesis to automatically locate faces of chimpanzees and gorillas in order to subsequently identify them. Thus, a more thorough review of the face detection framework by [113] is given in Section 4.2.1. In addition to real-time capable face detection in images and videos, the authors also propose methods to automatically distinguish between chimpanzees and gorillas. The first approach uses the detection scores of the applied face detection models for chimpanzees and gorillas. For the second method, a separate classification model was trained based on structure features only and applied to the detected face. Both techniques for species classification perform remarkably well with over 90% accuracy.

Although researchers use rigid object detectors mainly to localize the faces of animals, this class of object detectors has also been applied to detect whole animal bodies. Miranda *et al.* for instance use the face detection paradigm by Viola and Jones [99] for the visual detection of bumblebees [116]. Furthermore, the authors apply a discriminative tracking algorithm proposed by Gu and Tomasi [117] to improve the detection in video sequences and fill the gaps where detection has failed. Only one test sequence with restricted variety of different backgrounds and just one single individual was used to conduct experiments which is not enough for a thorough evaluation. Nevertheless, the obtained results are promising for a preliminary study. Although rigid object detectors can be used to detect insects such as bumblebees due to the limited number of their body appearances, the localization of deformable objects in natural scenes requires more sophisticated techniques because different body postures would result in different appearances of the same object.

Model-Based Detectors

To cope with the above mentioned challenges, researchers recently proposed model-based methods to reliably detect animals in visual footage. Different feature sets can be used to create models of an animal body such as appearance, texture, shape, color or the combination of those.

Stahl *et al.* reported a preliminary study on the combination of multiple 2D and 3D sensors to capture biometric data of farm animals under real-life conditions in [118]. More specifically the authors suggest to use a sophisticated hardware setup of two high-resolution monochrome cameras and one integrated color camera to reliably detect the heads of horses in a livestock farm. Due to the constrained environment at the feeding area and the fact that the horse's head points towards the camera, a simple foreground-background separation algorithm based on the depth information provided by the stereo cameras can be used to obtain a coarse location of the horse. To distinguish the animal's head from the rest of the body, an ellipse-like head model is subsequently created and fitted to the borders of the original head mask. This procedure

builds the basis to locate, measure, and identify the animal by fusing information of multiple cameras. By thoroughly evaluating the proposed framework, the system has proven to perform well. However, the application scenario as well as the controlled conditions in a livestock farm allows a multiple camera setup and a relatively simple geometrical approach to achieve an excellent detection performance. However, reliable detection of animals in more challenging natural environments requires more sophisticated solutions.

Remarkable ideas for a generalized unsupervised object tracker and detector were presented by Ramanan and Forsyth in [119, 120, 121]. In [119] the authors propose a technique to automatically build models of animals from a video sequence where no explicit supervisory information is necessary and no specific requirements regarding background or species are stipulated. Animal bodies are modeled as 2D kinematic chains of rectangular segments where the topology as well as the number of segments are not known a priori. In a first step, candidate segments are detected using a simple low-level local detector by convolving the image with a template which corresponds to parallel lines of arbitrary orientations and scales. This step alone results in many false positive detections. Therefore, resulting segments are clustered in a second step to identify body limbs that are coherent over time. After pruning away remaining segments that do not fit the motion model because the tracks are to short or move to fast, a coarse spatial model of the remaining segments can be assembled (see Figure 3.3). Because appearance models of animals are generated on-the-fly there are two ways to think about the proposed system: It can either be seen as a generalized tracking system that is capable of modeling objects while tracking them or as a source of an appearance model which can later be used to detect animals of the same species.

One drawback of this approach is that when building the rough shape model of an animal's body one has to be certain that within a given sequence only one animal is present and it is the only animal in the scene. Furthermore, the resulting appearance model is very much tuned to the specific species in the video sequence. Therefore, the same authors try to overcome these drawbacks by extending their work towards an object recognition framework to detect, localize, and recover the kinematic configuration of textured animals in real-world images [120, 121]. Ramanan *et al.* fuse the deformable shape models learned from videos and texture appearance models obtained from labeled sets of images in an unsupervised manner for that purpose. Although the detection results improve significantly compared to their previous work the detector is designed to only localize highly textured animals such as tigers, zebras, and giraffes. The approach would fail for a majority of camouflaged or non-textured animals such as elephants or great apes because detecting only vertical and horizontal lines to build the deformable model would result not only in false positive but more crucially false negative detections.

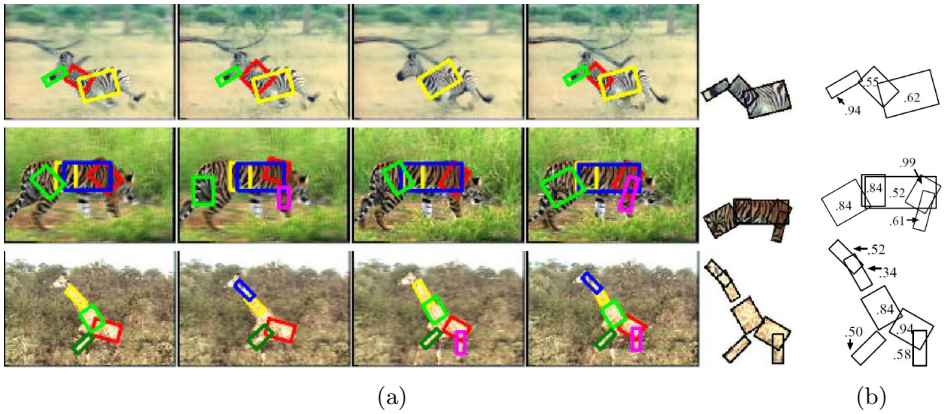


Figure 3.3.: ***Model-based detection of textured animals.*** (a) Each row shows four consecutive frames of moving textured animals. The detected segments are superimposed (left). The appearance models of the according animals are generated automatically (right). (b) Detection rates for every segment. The proposed algorithm builds representations of animals as collections of detected segments. Image source: [119, 120, 121]

More recently, Deformable Part-Based Models (DPM) were introduced by Felzenszwalb *et al.* [122] as generic object detectors. The proposed approach and its extensions regularly achieve state-of-the art results on a variety of object categories in international benchmarks such as the PASCAL VOC 2010 [123]. Whilst the majority of methods for object detection depict the object of interest as a whole, the main idea of DPMs is to represent objects as flexible configurations of feature-based part appearances. Support Vector Machines (SVM) are used to learn the set of appearance detectors and part alignments. While performing very well on some object categories, DPMs are known to be sensitive to occlusion and highly deformable object categories such as animals. Parkhi *et al.* [124] extend the ideas of DPM to distinctive part models (DisPM). They propose to initially utilize DPMs to detect distinctive regions such as the head of animals in the first phase. Secondly, the whole body of the animal is subsequently found by learning object specific features such as color or texture from the initially detected image region. After coarse foreground-background separation based on the learned low-level image features, graph cuts [125] are used for the final segmentation of the animal's body. Parkhi *et al.* apply their algorithm to detect cats and dogs in still images achieving results comparable to the state-of-the-art for other object classes. Furthermore the authors also claim that this technique can be used for a variety of other animal species as well.

Another study by Sandwell and Burghardt [126] uses three different models of DPMs to detect chimpanzee faces under difficult conditions such as far-off frontal poses or partial occlusion. Whilst the first and the second model are trained on the face region and an expanded version of the face region respectively, the third model integrates multiple spatially distributed DPMs.

Different from the approach presented by Ernst *et al.* in [113] the proposed algorithm is currently not real-time capable. However, the authors conclude that the reduced reliance on facial features alone and the combination of the three proposed models has led to a detector which is far less sensitive to non-frontal poses and more robust to less well resolved faces as well as partial occlusions.

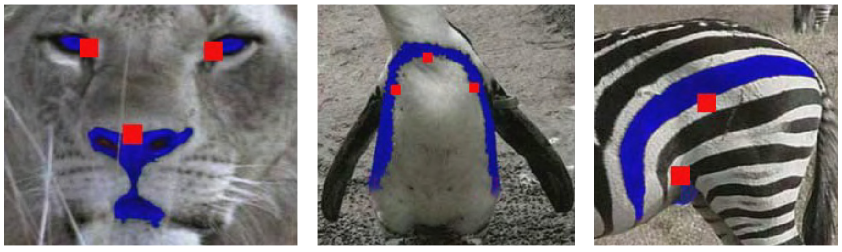
Local Keypoint Detectors

A large and growing body of literature investigates template matching, rigid object detectors and model-based approaches to detect animals in images or videos. However, they often perform poorly when detecting animals in their natural habitat due to a wide variety of postures, lighting conditions, and partial occlusion. Tiny object regions on the surface of an animal's body often exhibit less deformation than the entire organism [85]. Therefore, new approaches for fast, robust, and reliable detection and description of regions or local interest points such as the Harris Corner Detector [127], Scale Invariant Feature Transform (SIFT) [51] or Speeded Up Robust Features (SURF) [52] have been developed in the recent past. Inspired by the great success of local keypoint detectors and descriptors for object localization, matching, and categorization, these approaches have also been used for *Visual Animal Biometrics*. The applications include a variety of species, ranging from insect categorization [128] to turtle identification [129] and other coat patterned animals such as penguins and zebras [130, 131].

In 2010, de Zeeuw *et al.* proposed an approach based on SIFT matching for turtle detection and identification [129]. According to Zeeuw *et al.* leatherback sea turtles carry a so called *"pink spot"* on the dorsal surface of their head which is unique between individuals and can therefore be utilized for identification. Since the proposed algorithm is a semi-automatic approach, the first step is to manually cut the desired head region out of the original image. The extracted image patch is then compared with reference images using the basic SIFT matching approach which was originally proposed by Lowe *et al.* in [51]. More information about the identification scheme will be given in section 3.2.2. The evaluation on two challenging real-world datasets confirm the effectiveness and reliability of the algorithm proposed by Zeeuw *et al.* . However, manual interaction is still necessary to annotate the region of interest. Thus, designing a detection algorithm that automatically locates the pink spot on the turtle's forehead is highly desirable.

For the proposed approaches by Larios *et al.* [128] and de Zeeuw *et al.* [129] local keypoint detectors serve as a pre-processing step to locate stable points of interest for the subsequent extraction of discriminative information around each point. These descriptors are then used for individual identification, species classification or comparable tasks. Therefore, these approaches internally presume that the animal of interest is actually present in the processed image.

In [130], Burghardt exhibits keypoint detectors as initial detection stage to robustly detect coat patterned animals. Keypoint locations are initially stipulated by traditional corner detection algorithms based on the auto-correlation matrix of the input image over a small neighborhood. Since corner locations are defined by a significant signal change in all dimensions, the two eigenvalues of the auto-correlation matrix are analyzed to accurately detect corners in an image. More specifically, if both eigenvalues λ_1 and λ_2 have large positive values, then a corner is supposed to be found. This procedure is utilized in many corner detection algorithms such as the *Harris corner detector* [127] or the *Shi-Tomasi corner detector* [107]. In a subsequent step, the area around a detected point of interest is described by placing a neighborhood window around the keypoint. A class-specific point-surround classifier is learned by extending the Viola-Jones framework [99]. Instead of heuristically choosing the resolution of the neighborhood window, the dominant spatial frequency of the coat pattern is utilized to estimate a suitable window scale. Furthermore, a form of supervised *bootstrapping* is used to increase the robustness of the detector against false positive detections. Other modifications of the original implementation of the Viola-Jones framework refer to perspective constraints and dense belief maps, i.e. real-valued classification outputs instead of binary decisions. For details the interested reader is referred to [130]. Burghardt applies the proposed algorithm to detect frontal chests of Penguins, faces of lions, and hindquarters of zebras achieving results comparable to state-of-the-art human face detection algorithms. At a false positive rate of $4 \cdot 10^{-3}\%$ the detector achieves a detection rate of over 96%. However, false positive detections predominantly occur for highly cluttered background in natural environments, in challenging lighting conditions and hard shadows, as well as for cryptic resemblance due to groups of patterned animals imitating regional patterns of a single individual. Figure 3.4 illustrates a selection of typical reference points for different species.



(a) (b) (c)

Figure 3.4.: ***Detection of keypoint locations for different animal species.*** Species specific landmarks that are present for any individual within a population are marked in blue. A selection of the according reference points are visualized in red. (a) Lion face, (b) penguin chest, and (c) zebra hindquarter. Image source: [130]

Although local keypoint detectors are capable of robustly detecting and describing points of interest in certain scenarios, they disregard important global information such as body structure, spatial relationship between keypoints, and temporal information available in video sequences. Moreover, local keypoints can only be used for coat pattern animals or species with distinctive natural markings on fur or skin. However, many species do not carry obvious natural markings which can be used for detection.

Motion-Based Detectors

Although the above mentioned approaches achieve sufficient results under certain conditions, they lack in taking full advantage of the spatio-temporal information in video sequences to detect moving objects. Based on the assumption that even camouflaged animals can be detected by taking their movement in front of relatively stationary backgrounds into account, a considerable amount of literature has been published on motion-based detectors.

Especially the problem of detecting and tracking marine animals in underwater video has been tackled by many researchers [132, 133, 134, 135]. Either remotely operating underwater vehicles (*ROVs*) or live video feeds from stationary installed underwater cameras are used by biologists in order to perform marine ecological research within the *Fish4Knowledge*[1] project for instance. However, according to [135], state-of-the-art object detection algorithms often are unable to cope with difficulties which are present in unconstrained underwater environments such as cluttered and periodically moving background, permanent lighting changes, and the degrees of freedom of marine animal movements.

To overcome this issue Walther *et al.* [132] proposed a system capable of automatically detecting and tracking objects of interest in underwater video. Due to the fact that simple contrast-based detection algorithms are prone to a number of effects present in underwater video footage, such as non-uniform lighting conditions for instance, a background subtraction algorithm builds the first step of the framework. It was shown that even translucent foreground objects can be separated from the background sufficiently by subtracting the current frame from the mean image of at least 10 frames for every color channel separately. The actual object detection is subsequently applied using a saliency-based detection approach originally published in [136]. Furthermore, the authors found that oriented edges are a extremely useful features to detect marine animals and distinguish low-contrast translucent animals from organic debris. Once an object has been detected, Kalman filters [137, 138] are initiated to track the centroid of the detected objects. In a post-processing step, object tracks that are shorter than at least five consecutive frames are discarded as noise. A performance evaluation of the proposed system on

[1] `http://fish4knowledge.eu` Last visit: August 8^{th}, 2013

a single frame basis was conducted for two different data sets as well as a 10 minute video and achieved promising results with detection rates up to 80%. However, the question arises how many false positive detections were made by the system. Moreover, especially the evaluation on the detection and tracking performance is limited because the test set contains only a single 10 minute video. Two years later, the same authors extended their approach in [133] where a more thorough evaluation of object detection and tracking modules was conducted and the results of their previous paper were validated. Furthermore, the authors presented a technique to subsequently classify the detected objects into biological taxonomies. For each tracked object a feature vector based on Schmid invariants [139] is extracted and a Gaussian Mixture Model (GMM) is used for classifying the three most common classes. Although species classification was at a very early development stage at that time, promising results were achieved for the three examined animal categories.

A quantitative performance evaluation of object detection algorithms in underwater video footage can be found in [135]. The authors compared and thoroughly evaluated several state-of-the-art object detection algorithms for their application to detect moving objects in underwater settings. The dataset used for evaluation was taken from the publicly available *Fish4Knowledge* projects database. The performed experiments suggested that an algorithm called *Video Background Extraction* [140] outperformed the other approaches and therefore is most robust against the above mentioned challenges. However, the performance of all algorithms significantly decrease if typhoon events or storm is present in the video sequences.

Recently, Spampinato *et al.* proposed a framework for automatically analyzing fish behavior during typhoon events in real-life underwater environments [134]. In a first step, different texture features in combination with machine learning algorithms are used to detect typhoon or storm events in videos. Secondly, fish detection and tracking is performed. Similar to the work of [135] four algorithms have been implemented and compared against each other to detect fishes under extreme conditions. According to the authors, each approach performed fairly well were *Intrinsic Models* [141] achieved the best performance. However, often false positives were detected that had to be filtered out during post-processing. To deal with that problem, additional features such as color difference and difference of motion vectors at the object boundary as well as color homogeneity were extracted in a post-processing step and merged into a quality score. Only objects whose quality scores exceeded a pre-defined threshold were considered for further processing. Once the desired objects were detected, trajectories were extracted using a covariance-based tracking algorithm [142] which is known to adequately handle the typical challenges of tracking objects in underwater environments [143]. Within a last step, the extracted 2D trajectories as well as the object size, which indicates movement in the third dimension, were analyzed to evaluate movement patterns and behaviors of fish during typhoon

events. Each module of the proposed framework was evaluated on a sufficiently large data set of 257 video sequences of 10 different cameras and promising results were achieved by the overall system.

Besides approaches to detect marine animals in underwater environments, a growing body of publications has investigated motion-based detectors to localize mammals and birds in videos. In 2009, Wawerla *et al.* [144] reported an automated wildlife monitoring system called *BearCam* which was deployed near the arctic circle to detect grizzly bears in videos. The system was placed at a river site to monitor the animal's feeding behavior for four hours a day. To assist biologists with tedious annotation work, Wawerla *et al.* developed an algorithm for automatic detection of bear appearances in recorded video. The authors extend the shapelet features by Sabzmeydani and Mori [145] by additionally incorporating motion information from gathered video material. Shapelet features are a set of sophisticated mid-level features, originally developed to detect pedestrians in still images. They are constructed out of low-level gradient information using the AdaBoost learning algorithm [100]. In addition to simple gradient information as low-level descriptors, Wawerla *et al.* exploit background differences computed by taking the median over a sampled set of frames. Motion shapelet features are then constructed as a weighted combination of the previously extracted low-level gradient and background information within a specified sub-window using AdaBoost. In a third and final step, again AdaBoost is used to combine the information of different regions across the image and thus build the final classifier. A commonly used sliding window approach was used in the detection stage to localize the appearance of a bear in every frame. Extensive experiments proved the usefulness of the proposed algorithm which achieved adequate detection results for the task at hand. However, many false positive detections were found in highly textured regions and in areas with large amount of motion, e.g. at the banks of the river or regions of water. Moreover, because the detection is performed on every single frame the system was not real-time capable at that time. Object tracking algorithms could possibly speed up the performance while at the same time boost the detection accuracy of the system because non-moving objects might be removed in a post-processing step.

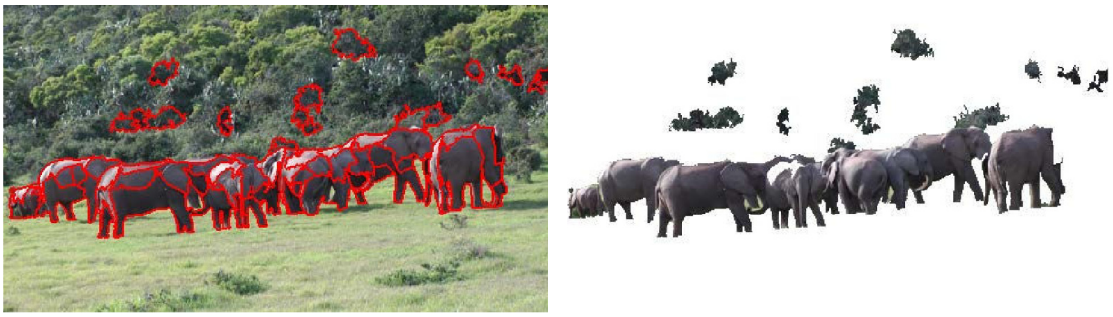
Song *et al.* [146] developed a robust autonomous system that assists ornithologists in observing and cataloging flying birds. Autonomous high-resolution video cameras were installed in the field which continuously scan the sky and automatically detect birds flying in the field of view. Video frames in which birds were detected are automatically send to the ornithologists for further processing. However, to save computational complexity and processing time only every fourth frame is scanned for birds using a non-parametric motion filtering technique proposed by Elgammal *et al.* in [147]. Similar to the background subtraction algorithm applied in [148], the method uses a Gaussian Mixture Model (GMM) to distinguish moving objects from constant background. Because the Gaussian distribution updates itself when a new sample comes in,

periodic movements by branches or trees can be characterized by the model. Because this technique alone would result in too many false positive detections due to cloud movements and other non-periodic motions, temporal difference filtering is used to estimate the velocity of detected objects in adjacent frames. Due to the fact that birds usually move a lot faster than clouds, false positive detections can be ruled out to a certain degree. Although the approach is rather simple, the authors found that during a long-term study of 310 days, where videos were captured continuously, 99.9953% of the data to be sent to the ornithologists could be removed by the proposed algorithm. However, because the system was designed to have a low false negative rate still 96% of that data was due to false positive detections. Moreover, the authors were unable to present a performance measure for the missed detections. Admittedly, Song *et al.* tried to measure the false negative rate using a two hour video. However, no bird was missed by the system in this single video file. Yet, because only every fourth frame is scanned thoroughly, it is very likely that birds might be missed in the long run.

In 2012, Khorrami *et al.* published a paper in which they described a system for the detection of multiple animal species in low frame-rate videos typically used by biologist to autonomously gather camera trap videos in wildlife environments [149]. The authors applied a recently developed technique for foreground-background separation called Robust Principal Component Analysis (RPCA) [150]. RPCA splits each frame of a video sequence in a low-rank matrix which contains pixels of the background and a sparse matrix representing the foreground activity of moving objects. An occurring animal is then isolated from the remaining foreground by calculating the local entropy for a small neighborhood around every pixel. While areas with similar intensities usually have low entropy values, abrupt changes due to edges caused by the boundary of the animal's body correspond to high entropy. Since this procedure still results in a relatively high number of false positive detections, the Large Displacement Optical Flow algorithm [151] detects large changes of velocity by incorporating motion information. The region with the highest amount of motion is considered to be the animal to be detected. Although the proposed method achieves promising results on a realistic dataset of ten different animal species, a high number of false positive detections occur in sequences with a high degree of background motion caused by rain and snow for instance. Another major drawback of the approach is that only one single individual at a time can be detected within a frame since only the candidate segment with the highest motion is considered to be the animal.

A system for automatic detection and tracking of elephants in unconstrained wildlife videos was recently proposed in [152]. Zeppelzauer *et al.* argue that current state of the art systems for animal detection and tracking often explicitly focus on highly textured animals. However, for animals with poorly textured skin such as elephants for instance other visual cues have to be be investigated. Also shape features would be impractical for the detection of animals due to pose

variation and partial occlusion in natural habitats. Therefore, the authors propose a method that learns a color model of elephants from few labeled training samples. In a first step, a mean-shift clustering algorithm [153] is used to extract spatial segments of the same color. Based on labeled training data, a Support Vector Machine (SVM) is trained to distinguish background color from foreground color in the LUV color space. However, as claimed by the authors, color alone is a weak and unreliable feature for elephant detection since many objects in natural environments have similar colors which leads to a unreasonably high number of false positive detections. Figure 3.5 shows the results of the detection based on color alone. As can be seen, many false positive detections occur in the background. To overcome this issue, the authors efficiently exploit temporal information to significantly reduce the number of false positives. Each initially detected segment is subject to a tracking algorithm based on the optical flow of the segment's pixels. Tracked segments are then joined into sets of coherent spatio-temporal candidate segments, i.e. segments belonging to the same objects are connected. Based on a number of extracted spatio-temporal features such as the tracking duration and changes of the segment's shape, the final decision of the appearance of an elephant is made. Since tracking of segments establishes temporal relationships over several frames, missing detections can be interpolated and tracking gaps are closed in a post-processing step.



(a) (b)

Figure 3.5.: ***Detection of a group of elephants in their natural environment using color information.*** (a) Regions classified as animals are marked in red. (b) Remaining segments after removal of textured background after detection. The few false positive detections are temporally not stable and can be removed later using consistency constraints of the spatio-temporal information in the video material. Image source: [152]

The proposed system was evaluated on a realistic dataset of elephant videos gathered under real-life conditions by biologists and achieved a high detection rate of 90% at a low false detection rate below 5%. Since the approach is claimed to be insensitive to pose variation, changing lighting conditions, and the number of individuals present in the video sequence, the authors did not have any further requirements for the tested video material.

However, researchers and gamekeepers often use infrared cameras in order to monitor animals during night. Thus, no color information is available in such recordings. Therefore, a different approach for initial object localization must be investigated in order to process gray-scale and infrared footage as well.

As presented in this section, a variety of different approaches exist to reliably detect animals in images and videos. Table 3.1 summarizes this section and compares the reviewed algorithms.

Category	Reference	Species	Description	Notes
Template Matching	Kastberger *et al.* [97]	Honey Bees (body)	(1) 3D stereoscopic imaging (2) Analyzing basic behavioral patterns (3) Detection and tracking of individuals	(1) Sophisticated hardware setup necessary (2) Template matching only achieves adequate results for species with non-deformable bodies
Rigid Object Detectors	Burghardt *et al.* [105]	Lions (faces)	(1) Viola-Jones AdaBoost cascade (2) Enlarged set of Haar-like features (3) Lukas-Kanade tracker of interest points (4) Basic locomotive behavior based on vertical head movement	Behavioral analysis could be improved by taking spatio-temporal interest points of the whole body into account
	Zhang *et al.* [108, 109]	Cat-like animals (faces)	(1) Jointly utilizing shape and texture using different alignment strategies (2) Haar-like features and gradients used as features (3) Fusing the results of separately applied detection models for final decision	(1) Extension of the algorithm to detect not only cats but also faces of cat-like animals (tigers, cheetahs, pandas) (2) Application restricted to cat-like animals only

Category	Reference	Species	Description	Notes
	Kozakaya *et al.* [110]	Cats (faces)	(1) Opposed to [108, 109] different feature sets are used to represent shape and texture simultaneously (2) Haar-like features and AdaBoost for candidate search (3) CoHOG and linear SVM for validation	(1) Outperforms approach by [108, 109] (2) Claimed to be more generic than [108, 109]
	Ernst *et al.* [113]	Great Apes (faces)	(1) Real-AdaBoost with multiple classification stages and increasing feature complexity (2) Gradient, structure, and census features used in different stages (3) Multi-resolution approach using image pyramids (4) Species classification based on detection confidences (5) Face tracking in videos using Kalman filters	(1) Suited for detection in still images as well as videos (2) Tracking of detected objects through video sequences (3) Lacks robustness to far-off frontal poses and severe occlusion (4) Used in this thesis as initial detection and tracking algorithm for subsequent identification
	Miranda *et al.* [116]	Bumblebees (body)	(1) Viola-Jones AdaBoost implementation in combination with Haar-like features (2) Body tracking using Gu-Tomasi tracker [117]	(1) Limited experimental verification (2) Treats insect bodies as rigid objects due to non-deformable bodies

Category	Reference	Species	Description	Notes
Model-Based Detectors	Stahl *et al.* [118]	Horses (head)	(1) Multiple 2D and 3D sensors (RGB and infrared cameras) (2) Coarse head localization using depth information (3) Fitting an ellipse-like head model for refinement	(1) Used for detection in livestock farms (2) Constrained application environment (head has to point towards the cameras) (3) Sophisticated hardware setup necessary
	Ramanan *et al.* [119, 120]	Various textured animals (body)	(1) Generalized unsupervised object detector and tracker (2) Localization of candidate segments using low-level texture detector (parallel lines) (3) Animal bodies are modeled by a 2D kinematic chain (4) clustering of segments to identify body limbs that are coherent over time	(1) No species-specific descriptors necessary for detection (2) Only one individual has to be present in the video (3) System is designed for highly textured animals only (4) Algorithm can be seen as a generalized object tracking system or as source for appearance model for detection
	Parkhi *et al.* [124]	Cats and dogs (body)	(1) Based on DPMs as initial distinctive region detectors (2) Detected region is used to learn species specific low-level features (3) Graph cuts [125] utilized for final body segmentation	(1) Approach outperforms DPMs for detection of animals (2) Performance could be further improved by incorporating further cues for learned appearance model
	Sandwell *et al.* [126]	Chimps (faces)	(1) Three different detection models (2) Integration of multiple spatially distributed DPMs	(1) Detection of non-frontal and occluded faces possible (2) Not real-time capable at that time

Category	Reference	Species	Description	Notes
Local Keypoint Detectors	de Zeeuw *et al.* [129]	Sea Turtles (head region)	(1) Detection of local interest points (SIFT [51]) at the "pink spot" located on turtle's head (2) Basic SIFT-matching [51] for identification (3) Verification based on affine transformation and gray-level pixel intensities	(1) Initial manual segmentation of area of interest necessary (2) SIFT features mainly used for identification
	Burghardt *et al.* [130, 131]	Penguins (body) Lions (face) Zebras (body)	(1) Initial detection of interest points based on eigenvalues of image auto-correlation matrix (see e.g. Harris corner detection [127]) (2) Learning of species-specific point-surround classifier by extension of Viola-Jones framework [99]	(1) High detection rates up to 96% (2) False positive detections mainly caused by highly cluttered background, hard shadows, and cryptic resemblance
Motion-Based Detectors	Walther *et al.* [132] Edington *et al.* [133]	Marine animals (body)	(1) Initial object localization using background subtraction (2) Validation by detection of salient regions and oriented edges (3) Kalman filters for object tracking (4) Post-processing to minimize number of false positives	(1) Promising detection results up to 80% (2) As shown in [135], more sophisticated motion-based object detection algorithms may further increase the performance

Category	Reference	Species	Description	Notes
	Spampinato *et al.* [134]	Marine animals (body)	(1) Comparison of several motion-based object detection algorithms (2) Difference of motion vectors and color features to further eliminate false positives in a post-processing step (3) Covariance-based tracking (4) Analysis of movement patterns during typhoon events	(1) Detection performance of all algorithms decrease significantly during storm or typhoon events (2) Behavioral analysis only for groups of fish, not for individuals
	Wawerla *et al.* [144]	Grizzly bears (body)	(1) Background subtraction based on frame differences (2) Extension of shapelet features [145] to motion shapelets in order to improve detection performance (3) Sliding window approach used for final detection	(1) High number of false positive detections in cluttered regions and segments with high motion (2) Detection is performed in every frame (3) Tracking algorithms could increase the performance in terms of speed and accuracy
	Song *et al.* [146]	Birds (body)	(1) Motion filtering based on GMM (2) Elimination of remaining false positives using temporal difference filtering	(1) Amount of data to be manually analyzed by ornithologists could be decreased significantly (2) Many false positive detections caused by background motion (3) Number of missed detections remains unclear

Category	Reference	Species	Description	Notes
	Khorrami *et al.* [149]	Multiple animal species (body)	(1) RPCA for foreground-background separation (2) Local entropy over small regions in combination with large displacement optical flow for refinement	(1) False positive detections caused by high degree of motion due to rain or snow (2) No restriction to a certain species (3) Only a single animal in each frame can be detected
	Zeppelzauer *et al.* [152]	Elephants (body)	(1) Learned color model and SVM for foreground-background color classification (2) Optical-flow based tracking (3) Post-processing to decrease number of false positives	(1) Algorithm is claimed to be insensitive to pose variation, lighting, and number of individuals (2) Algorithm would fail for gray-scale and infrared video footage

Table 3.1.: ***Overview of state-of-the-art algorithms for animal detection in images and video footage.***

Tracking and Behavioral Analysis of Animals

Besides the pure detection of animals in image and video footage, the field of *Visual Animal Biometrics* also offers the possibility to automatically analyze specific behavioral patterns of certain species. Although usually bound to controlled environments, sophisticated spatio-temporal features combined with state-of-the-art machine learning approaches can help to automatically recognize behaviors of mice for instance [154, 155]. As done in [148], Serre *et al.* use the implementation of Stauffer and Grimmson's GMM-based background subtraction algorithm [156] to initially detect captive mice. After detection, Serre *et al.* provide a neuro-biologically motivated computer vision algorithm for automated analysis of complex mouse behaviors. The algorithm is based on earlier work by the same authors [155] which uses a hierarchy of position-invariant spatio-temporal feature detectors with increasing complexity. Even for minute movements which are typical for mice, a combination of gradient based features, optical-flow based descriptors [157], and space-time oriented Gabor filters [158] as well as a multi-class Support Vector Machine (SVM) for classification achieves excellent recognition results on eight different behaviors such as drinking, eating, grooming, walking, etc. Moreover, several studies have

been conducted to analyze the behavior of social insects in captive environments, more specifically ants [159] and bees [160, 161]. The basic assumption of most state-of-the-art tracking algorithms is that targets maintain their behavior, e.g. velocity and direction, before and after interaction with other objects. However, when it comes to tracking of animals, this assumption does not hold because animals often do not act independently. To overcome this limitation, a framework for tracking multiple interacting ants has been proposed by [159] in 2004. Kahn *et al.* extended a joint particle filter by including a more sophisticated motion model that reflects additional complexity of the behavior of multiple interacting targets. Also the study of communication forms in bees has received much attention in recent years [160, 161]. Oh *et al.* [160] introduced a Parametric Switching Linear Dynamic System (P-SLDS), an extension of the method proposed in [162], for the description of complex temporal patterns. They applied this algorithm to automatically label and interpret characteristic movements of honey bees. They use the particle filter based tracking algorithm proposed in [163] to extract the trajectories of tracked bees. An Expectation-Maximization (EM) algorithm [164] is used to learn the necessary P-SLDS parameters from the obtained trajectories to distinguish between three movement patterns of the so-called *waggle dance*, a form of bee communication which takes place in hives. Inspired by this work, Veeraraghavan *et al.* published their work on shape and behavior encoded tracking of bee dances in [161]. In contrast to the work by Oh *et al.* [160], the position and orientation of bees are explicitly parametrized using a shape model. Motion behaviors are classified using a hierarchical Markov Model similar to those used in [165]. Furthermore, while Oh *et al.* treat tracking and motion classification as two completely independent processes the approach by Veeraraghavan *et al.* enables simultaneous tracking and behavioral analysis. This aspect enhances the tracking accuracy while at the same time enables accurate interpretation of behaviors. Other attempts to automatically analyze the behavior of animals include mammals [166, 105], birds [167, 168] as well as fish [134]. While the three latter studies only consider the behavior of flocks of birds or groups of fish rather than the motion activity of an individual, Gibson *et al.* analyzed the quadruped gait of individual animals in [166]. A sparse set of tracked points provided motion signatures that were used to determine the presence of an animal and subsequently classify basic movement patterns. A rather simplistic approach for behavioral interpretation was used by Burghardt *et al.* in [105], where the vertical trajectories of detected and tracked lion faces were used to classify basic locomotive activities such as walking, trotting, or standing.

Although the proposed algorithms provide first approaches for behavioral recognition of animals, the development and thorough evaluation of algorithms capable of accurately and robustly recognizing more complex behaviors under natural real-world conditions remain to be implemented.

3.2.2. Animal Identification and Species Recognition

After successful localization (either manually or automatically) of animals in images or videos one of the most common animal biometric applications is the identification of individuals. According to [85] identification is the process of retrieving the identity of an unknown test sample by matching visually unique features to a set of known biometric profiles. Often, invasive methods for individual identification of animals such as branding, ear tags, microchip implants or leg rings, have been used and are still being used for animal monitoring. However, as stated earlier, such intrusive techniques generally involve severe interaction with animals, including capture and handling, which often leads to modified behavior and physiology [169] as well as affection of activities and relationships to other animals [170]. Moreover, it is documented in several studies that artificial tags for identification can cause permanent pain [91, 92] and might even affect survival of marked animals [171, 29, 30, 31].

As a consequence, researchers started to develop non-invasive techniques for animal identification. As stated by Kühl and Burghardt in [85], early methodological work for identification of animals based on distinctive body markings dates back to the mid-1900s. A large and still increasing number of long-term studies of mammals have shown that natural markings can be utilized to identify individuals of several species including zebras [172], giraffes [173], lions [174] and many more. Especially within the marine animal observation community visual comparison of naturally occurring notches in flukes or dorsal fins have been used for individual identification [175, 176, 177, 178]. Although these methodologies paved the way towards objective identification by means of phenotypic appearance, manual comparison of visual markings were error prone, extremely time consuming and therefore not applicable for large datasets. Due to the increasing number of keen digital recording devices and the great progress in the field of automated pattern recognition over the past two decades, recently a lot of effort has been put into the development of automatic techniques for animal identification. Nowadays, a number of innovative (semi-)automatic software tools such as *Wild-ID* [179], *HotSpotter* [180], *Photo-ID* [181, 182], *MantaMatcher* [183], *StripeSpotter* [184], or *Sloop* [185] are available to assist researchers with tedious annotation work. While some methods are species specific e.g. [184, 183], others are not limited to a certain species and are able to recognize patterned animals in general, e.g. [179, 181, 182, 185].

Although primarily developed to identify free-living animals in natural habitats, recently non-invasive methods for cattle recognition has entered livestock breeding. For instance methods for automatic analysis of muzzle prints [186, 187, 188] or iris pattern [189, 190] have recently been proposed to replace more traditional methods for cattle identification to maintain a livestock. A thorough literature review and discussion about the use and abuse of traditional methods and biometric techniques for animal identification in livestock farms can be found in [191].

This chapter reviews the literature published in the field of non-invasive animal identification with a primary focus on recognition in natural settings and wildlife habitats. Techniques for computer-assisted animal identification can be classified into three main categories: *Contour-Based* methods, *Pattern-Based* approaches, and techniques based on *Face Recognition*.

Another interesting application of *Animal Biometrics* is automatic species classification of the previously detected animal. Such approaches have recently gained huge interest within the computer vision community due to their wide applicability in biodiversity monitoring. At the end of this chapter a comprehensive overview of existing methods for automatic species classification is given.

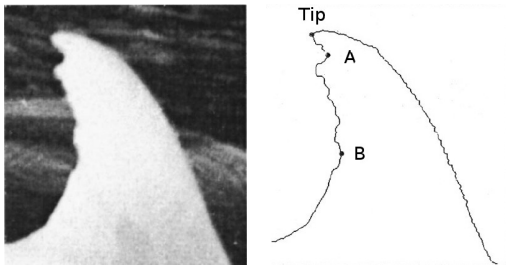
Contour-Based Methods

Many species, especially marine mammals such as dolphins, sea lions or whales, can be uniquely identified by markings on their flukes or dorsal fins [192, 193]. Such notches are often caused by bites or other interactions with conspecifics, predators or humans and once acquired are often permanent over time [192, 194]. Thus, the presence of such marks for reliable photo identification has been used by marine biologist for over two decades. However, manual inspection is time consuming, subjective, cumbersome, and becomes impractical for large datasets. Therefore, a considerable amount of literature has been published on contour-based computer vision algorithms to automate this process.

One of the first studies of computer-assisted identification of bottlenosed dolphins in images was published by Hillman *et al.* in 1998 [195]. At that time user-interaction was required to interactively label the region of interest, align the images for comparison, and identify the tip of the fin. The authors proposed two techniques to represent the fin edge. The first is based on affine-invariant Fourier descriptors originally developed by [196]. For the second method, a smooth approximation of the curvature was generated by fitting a fourth-degree polynomial to the boundary of the dorsal fin. Normalized peak positions with respect to the tip of the dorsal fin as well as the width and height of the localized peaks serve as descriptors. However, a conclusive evaluation to compare the proposed features against each other and to show their effectiveness on suitable datasets was not performed by the authors.

In 1999, Kreho *et al.* [197] proposed a computer-assisted method to extract the so called dorsal ratio, a measurement commonly used by marine biologist for manual identification of dolphins. The dorsal ratio is defined as the ratio of distances between the tip of the fin to the two largest notches [193] (see Figure 3.6). The result of the Laplacian-Of-Gaussian (LoG) edge detector [198] serves as a pre-processing step for the subsequent active contour model [199], a well known technique for smooth curvature fitting and conformation of contours.

Although less user-interaction is required compared to the method proposed by [195], the start and end point of the fin's edge has to be selected manually to speed up processing time and to guarantee convergence of the algorithm at different scales. Once the spline curve has been fitted and an appropriate scale has been found, the dorsal ratio can be calculated by taking the most prominent maximum (the tip of the dorsal fin) and minima (the two largest notches) of the curvature function into account. Figure 3.6(a) shows a typical photo of a dolphin's fin with notch patterns on the trailing edge, while 3.6(b) illustrates the calculation of the dorsal ratio based on the extracted curvature.



(a) (b)

Figure 3.6.: ***A typical photograph of a dolphin's dorsal fin before and after edge detection.*** (a) The prominent notches on the trailing edge are often used by biologists to distinguish between individuals. (b) After edge detection and curvature fitting, the dorsal ratio (DR) is defined as the ratio of distances between the tip of the fin and the two most prominent notches: DR= (A to B)/(B to Tip). Figure adapted from [200].

On a dataset of 296 images of 94 individual the authors showed that the proposed algorithm is comparable and even slightly better that manual analysis of the images by non-experts, while the processing time is drastically reduced up to a factor of 15. However, the authors also argue that for a reliable identification the dorsal ratio alone is not enough and additional features have to be taken into account. Therefore, the proposed method is only suitable to narrow down the search by presenting a ranked list of possible identities.

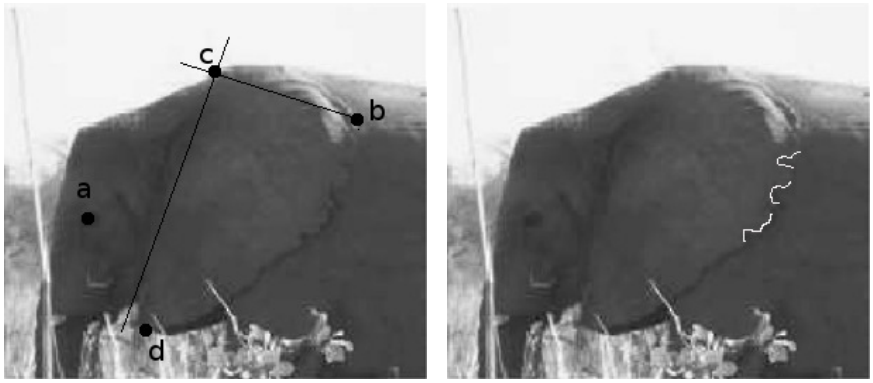
An extension of the above mentioned methods was presented in 2000 by Araabi *et al.* in [200]. As done in the the study of Kreho *et al.* [197], the LoG edge detector as well as the subsequent active contour model is used to get a smooth approximation of the fin's curvature. In addition to previous algoritms however, Araabi *et al.* propose to extract additional attributes from the trailing edge of the dorsal fin. A curvature function can be seen as a sequence of positive and negative areas separated by zero crossings of the curvature. These primitives can then be labeled with low-level rotation and translation invariant attributes such as *area*, *length*, *width* and *height*. Hence, a representation of the approximated curvature function can be extracted and compared with descriptors in the dataset. Therefore, unlike the dorsal ratio used in [197], the

proposed feature vector reflects not only the two most prominent structures but all significant curvature elements that are useful for identification. The proposed algorithm shows superior performance on a dataset of 624 images of 164 manually identified individuals over previously published algorithms and outperform the approaches proposed in [197] and [195]. Although the developed computer-assisted method for dolphin identification shows encouraging results and can therefore support marine mammalogists with tedious routine work, a significant amount of user-interaction is still required. Therefore, the average processing time to get a matched individual is with about 5 minutes per image relatively long. Moreover, although the proposed curvature attributes are invariant against rotation and translation, they cannot suitably cope with affine transformations which are often present in image datasets gathered under realistic conditions.

To address the problems resulting from out-of-plane rotations, Gope *et al.* proposed an affine invariant curvature matching strategy in [201]. The extraction of the curve of an animal's body is achieved by first applying an interactive semi-automatic edge detection paradigm proposed in [202]. Subsequently, a spline curve is fitted to the detected edge in order to approximate the curvature. A matching strategy between a query and a database curvature is then applied by first using an appropriate affine transformation such that both curves overlap as much as possible. The mismatch area between both aligned curves serves as a distance measure which is exploited to calculate a ranked list of possible identities. The authors evaluate their algorithm on three different databases of marine mammals, namely sea lions, dolphins, and gray whales. The performance statistics suggest that the proposed method is capable of considerably reducing the number of images to be searched manually for all three datasets. However, instead of benchmarking against previously developed identification tools (e.g. [195, 197, 200]) the authors compare their algorithm to other affine invariant curvature matching strategies. Therefore, the benefit of the proposed algorithm over traditional software tools for marine mammal identification remains unclear.

Although contour-based recognition methods were mainly developed for marine mammals, Ardovini *et al.* adopted these ideas for elephant photo identification. Based on the assumption that elephants carry individually characteristic nicks in their ears, the authors proposed a computer-assisted semi-automatic tool for elephant identification in [203]. Unlike previously published algorithms for contour-based identification, the proposed method is robust to partial occlusion because identification is based on matching non-connected curve sets. However, a significant amount of user-interaction is still required. First, a set of reference points is needed in order to approximately define the elephant's head position, orientation and scale (see Figure 3.7(a)). After applying the canny-edge detection algorithm [204], a second user feedback is required to find the start and end point of for each nick in the elephant's ear. Figure 3.7(b)

shows the detected nicks that are subsequently used for identification on an example image.



(a) (b)

Figure 3.7.: ***Example profile image of an elephant's head.*** (a) The user has to manually select a number of reference points (a-d) in order to define the elephants head location, orientation, and scale. (b) Detected ear nicks are marked as white lines. The proposed method is robust to partial occlusion since non-connected curve sets are matched. Figure adapted from [203].

The actual matching strategy is based on two main steps. First, global matching achieves a subset of curves having consistent locations with respect to the keypoints selected by the user. Secondly, each valid global match is locally verified by pair-wise comparison of their shapes. For testing and evaluation the authors used a set of 532 images of 268 different individuals achieving promising results. The probability that the correct individual is among the first 5 photos proposed by the system is about 75%.

Contour-based methods have been used successfully by biologists and marine mammalogists to identify dolphins, whales, sea lions, and even elephants based on individually characteristic notches in flippers, flukes, dorsal fins or ears, respectively. Unfortunately, due to cluttered background, difficult lighting conditions and other challenges in real-world environments often a significant amount of user-interaction is required in order to achieve sufficient accuracies. Although most authors state that distinctive edges on animal bodies are permanent and stable over time, other studies have revealed that fluke and dorsal fin features can undergo severe changes over time [205]. Moreover, contour-based methods can only be applied to a small number of animal species with individual distinctive edge features on their bodies. However, many animals carry prominent visual markings, commonly known as *coat patterns*, on areas of their body surface. These markings are often unique and complex enough to distinguish individuals of a population [206].

Pattern-Based Methods

Pattern-based methods for computer-assisted animal identification exploit naturally evolved individually distinctive markings on fur or skin of coat patterned animals to implement automatic procedures. The purpose of these methods is to extract a compact and unique representation of the animal's body texture which can be compared to extracted descriptors of known individuals in a database. Pattern-based identification techniques can be subdivided into dense pattern-based identification schemes and sparse texture comparisons [130].

Dense Pattern-Based Methods Dense pattern-based identification algorithms aim at extracting individually distinctive image features based on a global representation of the animal's body texture. Dense features are exploited to classify the identity of animals within a population of conspecifics. Due to the fact that most animal bodies are flexible entities, i.e. the shape of the bodies and imprinted coat patterns can change their appearance over time depending on the animal's posture, dense sets of features are usually extracted at a particular region of interest (ROI) which can reduce possible deformation of texture. However, often an additional alignment step is necessary in order to achieve comparability of textures over the dataset and thus decrease intra-class variability.

One of the pioneering works of dense pattern-based methods for animal identification was published in 1990 by Hiby and Lovell [181]. The system exploits pelage markings on the head and neck of gray seals for identification. To compensate pose variations, a pre-calculated 3D surface model of a gray seal's head is projected onto the image and fitted to the outline of the animal. Manual initialization of three keypoints is necessary in order to accurately fit the model to each new photograph. After the model is fitted, the user is asked again to adjust the model in order to refine the result. The numerical descriptors within a pattern cell, a subregion of the fitted 3D model, are simple gray-scale intensities. The similarity between different identifier arrays is defined as the correlation coefficient between corresponding elements. The same 3D matching system was successfully used and tested by Kelly in [207] to identify Serengeti cheetahs in wildlife photographs. A major drawback of the proposed system is the simplicity of the features used for identification. Simple grey-scale pixel intensities are highly susceptible to different lighting conditions and lack discriminative power to reliably distinguish individual animals. Therefore, later these ideas were extended by Hiby *et al.* to identify tigers by means of distinctive stripe patterns on the animal's fur [182]. Again a 3D model is fitted to the deformable body based on manually defined keypoints, which allows comparing coat patterns across changing camera angles and postures (see Figure 3.8).

The results of two different complementary feature sets are combined for identification. The posterior probability that the patterns are from the same individual given the scores of both

Figure 3.8.: ***An example image of a camera trap showing a tiger.*** A 3D model is fitted based on manually defined keypoints. The red and blue points denote the upper and the lower bound of the animal's body, while the yellow spots show the landmark points defined by the user in order to place the model into the scene. The texture within a region of interest defined by the fitted 3D model is then aligned and features are extracted (lower right corner of the image). Image source: [182]

matching strategies are utilized for decision-level fusion. As in the previous studies, a ranked list of possible identities has to be inspected manually by the operator. By now the proposed identification tool has been extended to several species including sharks, zebras, lynx, and many more.[2] Although the Hiby-Lovell system has proven itself successful for a variety of different species [207, 208], one drawback of the system is the requirement of intensive user interaction to accurately fit the surface model to new images. This procedure is not only tedious and time consuming for large datasets but could also introduce subjectivity when different users are involved.

Although the majority of techniques available for marine mammal identification are curvature-based methods, a growing amount of literature investigates the feasibility of natural patterns on the fluke of whales to distinguish between individuals. Black and white patches on the flukes of humpback and gray whales constitute unique features that are useful for identification. One of the first approaches for dense-pattern based identification of whales was presented by Kehtarnavaz *et al.* in [209]. A semi-automatic edge detection algorithm developed by [202] was used as a pre-processing step to segment the fluke area from the background. Otsu's method for optimal thresholding [210] was subsequently adopted to obtain a binary image of dark and light patches. Afterwards, Affine Moment Invariants (AMI), first derived by Flusser and Suk [211], are extracted from the binary image patches and used as individual representation. Since these features are robust to affine transformations such as translation, rotation, scaling, and skewing an additional alignment step is not necessary in this case. The final matching is finally done

[2] `http://www.conservationresearch.co.uk/`. Last visit: December 10^{th}, 2013

by a minimal euclidean distance measure between the unknown test sample and all samples in the gallery. On a dataset of 38 individuals it has been shown that the proposed technique outperforms the contour-based method by Araabi *et al.* in [200]. An improved version of the proposed dense pattern-based algorithm was presented in 2004 by Gope *et al.* in [212]. The feature extraction method based on Affine Moment Invariants was replaced by so called Zernike Moment Invariants (ZMI), a set of orthogonal polynomials defined on the unit disc which have been shown to perform well in image recognition tasks [213]. Furthermore, the authors propose to use Linear Discriminant Analysis (LDA) to transform the extracted feature vectors into a smaller, more discriminative subspace. Similar ideas for whale identification based on dark and light patches on their flukes have been discussed by Ranguelova *et al.* in [214]. However, their approach differs in the utilized methodology. Instead of extracting affine invariant image descriptors, the authors of [214] propose to construct a grid that is semi-automatically fitted to the fluke based on three manually selected keypoints. This grid divides the fluke into small but affinely invariant regions. After applying Otsu's optimal thresholding algorithm [210] to obtain a binary image, the relative contribution of black and white patches in every grid cell serves as feature. After concatenating the descriptor of every region, the final feature vector for the query fluke can be compared with the extracted features in the database by using a simple Euclidean distance based nearest-neighbor classifier. The authors extended this approach in [215], where salient patterns are characterized by fitting an ellipse to every detected patch. Each pattern is therefore represented by the four extrema points of the fitted ellipse and the position of the grid cell it is located in. By combining the approaches presented in [214] and [215], the system's robustness to different image quality levels could be improved significantly.

Although the evaluation results for all three studies indicate that the number of images which have to be manually searched by an expert can be significantly reduced, still a significant amount of user-interaction is required in order to reliably segment the fluke from the background. Moreover, for low resolution images and photographs with insufficient contrast, small patches on the whale's flukes are often missed during the pattern detection stage which hampers correct classification in some cases.

Due to their obvious visual stripe patterns which are comparable to a human fingerprints, computer vision algorithms have also been developed to identify individual zebras. One of the first approaches for non-invasive automated recognition of zebras in field photographs was proposed by Krijger *et al.* in [216, 217]. Based on six manually defined keypoints, a number of pre-processing steps including de-interlacing, smoothing, and block-wise adaptive binarization within a region of interest is performed in order to enhance the image quality. Subsequently, a method called sequential thinning [218] was applied to obtain a skeletonized version of the stripe patterns. Krijger *et al.* compared several different matching strategies ranging from simple image

overlay techniques to methods commonly used for biometric identification of human fingerprints. However, stripe patterns of zebras contain less information than a human fingerprint. Therefore, a number of empirical measures such as curvature score, slope score, as well as intersection point correlation were recorded for comparative analysis. In 2011, Lahiri *et al.* presented another system for individual zebra identification called *StripeSpotter*[3] in [184]. As in the study by Krijger *et al.* , the procedure of the proposed method starts with a manual selection of a region of interest (see Figure 3.9(a)). After binarization a simple but effective animal coat feature called *StripeCode* is extracted. As claimed by the authors, this feature is not only easy to implement, but is also invariant to scale, exposure, partial occlusion, and mild skew. A dynamic programming algorithm based on the work of [219] is utilized to compare the descriptors of two images. Figure 3.9(b) shows the extracted *StripeCode* of an example image as well as the comparison of two extracted descriptors of the same zebra.

(a) (b)

Figure 3.9.: ***A field photograph of zebras and the extracted*** **StripeCodes.** (a) First, the user has to manually define a ROI before *StripeSpotter* can extract unique descriptors for matching. (b) Red blocks of the left *StripeCode* are matched to blocks of the right *StripeCode* extracted from two images of the same zebra. Brighter colors indicate low matching costs while dark colors represent blocks with a high cost. Image source: [184]

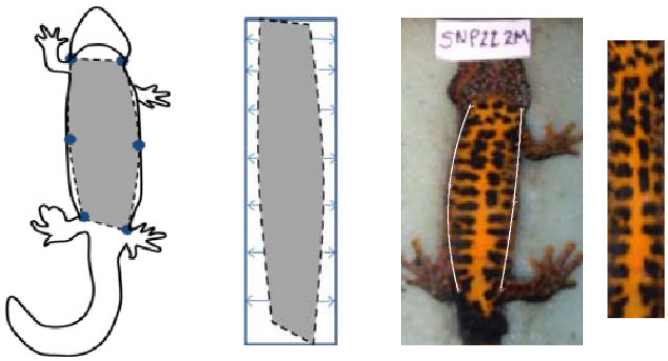
The interested reader is referred to [184] for details. Although the evaluation results show the effectiveness of the algorithms proposed by [216] and [184], manual ROI selection can be extremely tedious for large datasets. However, as shown by Burghardt in [131, 130], this process can be further automatized using local keypoint detection methods (see section 3.2.1 for details).

Not only mammals carry individually characteristic natural markings on their bodies. Also a variety of amphibian species such as salamanders, newts, and frogs can be recognized by exploiting unique visual traits on their skin for computer-aided identification. Because recog-

[3] `http://code.google.com/p/stripespotter/`. Last visit: December 10^{th}, 2013

nizing individual amphibians is an important step in assessing migratory patterns in ecological studies, Ravela and Gamble suggested one of the first approaches for automated identification of marbled salamanders in [220, 221]. To reliably segment the body of salamanders, they are photographed at field sites by placing them in a custom enclosure with diffuse lighting and constant background during acquisition. Due to the flexible nature of salamanders, the segmented body has to be virtually straightened in order to align the patterns throughout the dataset. This is done by manually annotating a series of points along the dorsal midline of the body. By interpolating between those points, a smooth curve that outlines the shape of the body can be fitted and then warped onto a straight line. Secondly, the user is asked to select four pre-defined keypoints on the straightened salamander to approximate the mid-section of the body. Inspired by ideas of [222, 223, 224, 139] the authors propose to represent the local structure of the image by applying a set of multi-scale Gaussian Derivative Filters (MGDFs). The extracted features are then composed into shape and orientation histograms over several scales withing overlapping windows along the length of the salamander in order to get a compact multi-scale appearance representation. The normalized cross-covariance between the obtained query histogram and each database histogram vector is utilized as similarity measure. On a self-established database of 370 images of 69 individuals impressive results were achieved, where approximately 91% of the query images were ranked at the top 5 individuals proposed by the system. These results were confirmed and even improved in a follow-up study by the same authors on a larger dataset containing 1008 images of 366 individuals. These ideas were recently incorporated together with other state-of-the-art computer vision algorithms into an individual identification system based on interactive image processing and matching with relevance feedback from crowdsourcing called *MIT SLOOP* [225, 185]. This image retrieval system for *Visual Animal Biometrics* is therefore one of the first collaborative community-based frameworks that can be used by biologists for conservation planning. As yet, it has been applied to a variety of different patterned species including skinks, whales, and marbled salamanders.

In 2011, a preliminary study for computer-aided identification of newts was presented by Hoque *et al.* in [226]. The belly of newts displays distinctive patterns of dark spots which can be used for identification. Again, a significant amount of user interaction is required to align the textures on highly deformable bodies. Therefore, the operator has to manually select six keypoints on the outline of the animal's body. A cubic spline curve is subsequently fitted to these point to selected the region of interest which is then stretched to fit a predefined rectangle to align the texture. Figure 3.10 illustrates the interactive alignment process.



(a) (b) (c) (d)

Figure 3.10.: ***Texture alignment for the identification of newts.*** (a) Six keypoints on the outline of the newt's body have to be selected by the operator. (b) The manually defined ROI is subsequently stretched to fit a predefined rectangle of fixed size. (c) A spline curve is fitted to the defined keypoints and therefore defines the ROI. (d) The aligned texture of a newt's deformable body. Image source: [226, 227]

In order to match the images of two newts, the corresponding ROIs are compared using the correlation coefficient of the vectorized gray-scale intensities. Since the comparison of simple gray-scale pixel values is extremely susceptible to numerous extrinsic factors such as lighting, noise or image artifacts, the approach was extended by Azhar *et al.* in [227]. The authors propose to utilize multi-radii normalized Local Binary Patterns (LBP) histograms [50] as compact and meaningful representation of newt patterns. LBP histograms are well-known features that have been proven to be robust descriptors for texture representation and classification due to their discriminative power and robustness to various image variations. As done in the original work by [50], the chi-squared distance metric between two histogram sequences is used for matching. As expected, the proposed approach outperforms the system proposed by [226]. However, although images were acquired in relatively controlled environments, still a lot of manual preprocessing is required by both systems in order to align the body textures and achieve adequate results on real-world datasets.

A completely autonomous system for frog identification was recently proposed by Cannavò *et al.* in [228]. Photographed in a controlled environment with constant background, a particular part of the frog's body is first detected using Otsu's binary thresholding scheme [210] in combination with a number of standard morphological operations. The second order statistics of the largest resulting area - which is supposed to be the frog's body - are used to calculate the orientation of the frog. A normalized region on the frog's backside is subsequently chosen for feature extraction since this area is less effected by body deformations. Different low-level features are exploited to get a characteristic representation of the body's texture which can later

be used for identification. The experiments indicate that a low-level texture descriptor called granulometry [229], a measure of size distributions of grains in binary images, is well suited to identify frogs based on their body patterns. However, experiments were performed on a dataset consisting of 60 different individuals with only a single picture per frog. To get a sufficiently large dataset a number of different modifications have been applied to every image. Although the proposed algorithm shows excellent results with over 96% accuracy, experiments were performed on an artificial dataset which reflects the conditions of natural settings only to a certain extend.

The success of the above envisaged automated dense pattern-based animal identification techniques shows the potential of naturally occurring markings for the recognition of individuals. However, using dense texture descriptors on unaligned data contents a high risk of false classifications due to misinterpretation of patterns before matching. Therefore, due to the challenging task of automatically aligning textures gathered in real-world environments almost all dense pattern-based approaches rely at least partly on human interaction. Although researches have been making great efforts to reduce the amount of human interaction, especially for large datasets manually selecting a number of reference points for every image in a database is not only time consuming and tedious but also prone to errors due to tiresome routine work. To overcome these limitations, a significant amount of literature has been published on sparse pattern-based methods for animal identification. Based on the fact that sparse, localized features on the two dimensional surface of patterned animals can provide a great wealth of information, descriptors are calculated only around a restricted area of automatically detected keypoints to construct biometric profiles. Thus, an additional alignment step is often not necessary for sparse pattern-based identification techniques.

Sparse Pattern-Based Methods Inspired from the success of local keypoint methods for object localization, recognition, and matching, a large amount of literature has been published that adapt such sparse pattern-based techniques for animal identification. The main idea of sparse pattern-based computer vision algorithms such as Scale-Invariant Feature Transform (SIFT) [51] as well as its extensions and variants [52, 230, 231, 232, 233] can be summarized as follows:

1. Detect distinctive keypoints in an image using an interest operator in scale-space;
2. Represent the image patch surrounding each interest point using a local descriptor that is invariant to expected transformations and
3. Store each keypoint descriptor for efficient matching of interest points across images.

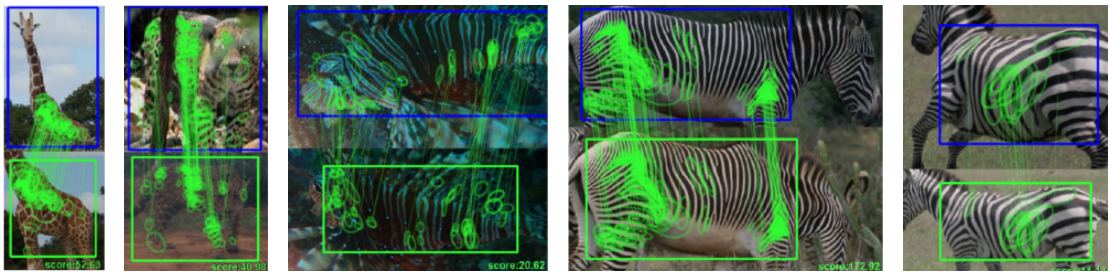
One advantage of sparse pattern-based methods is their invariance to uniform scaling, orientation, and partly even small affine transformations. Therefore, the necessity of texture alignment - which is a crucial pre-processing step for dense pattern-based methods - does not apply if keypoint descriptors are used for matching.

One of the first studies that adopted SIFT or SIFT-like features to automatically distinguish between different individuals was published by Pauwels *et al.* in 2008 [234]. The authors apply the SIFT detection, description, and matching procedure proposed by Lowe *et al.* in [51] to the so called "pink spot" on the head of leatherback sea turtles. This area carries individually distinctive markings which are commonly used by biologists to identify individuals by eye. Later Zeeuw *et al.* extended these ideas in [129]. In addition to pure SIFT matching, Zeeuw *et al.* proposed to refine the results by checking the cross-correlation coefficient of a number high-contrast regions on the query image and an transformed version of a possible match. Although the results confirm the effectiveness and accuracy of the proposed system, the "pink spot" has to be selected manually for every image which might be tedious and time consuming for large datasets.

Three of the most widely acclaimed tools available for sparse pattern-based animal identification are *Wild-ID* [179], *HotSpotter* [180], and *MantaMatcher* [183]. In 2012, Bolger *et al.* developed a stand-alone cross-platform pattern extraction and matching application for mark-recapture analysis called *Wild-ID*[4] [179]. The system is based on an open-source Java implementation of Lowe's SIFT feature detector and descriptor [51]. Due to the robustness of SIFT against scale changes, rotation, and different lighting conditions no particular pre-processing step is necessary. However, since a brute-force candidate matching of all detected features of the query image and all training images is employed, the authors state that cropping the region of interest beforehand is highly recommended in order to minimize the influence of background clutter. To further improve the performance of the original SIFT matching approach, Bolger *et al.* additionally exploit a modified version of the RANSAC algorithm [235] to find a geometrically self-consistent subset of the keypoints previously matched. Three candidate matches are randomly chosen from the test image. By searching for a corresponding triangle with similar geometry in the reference image, matched keypoints can be further verified. The proposed software package was evaluated on a self-established dataset of 50 giraffes by presenting a ranked list of the 20 most probable individuals. Measured error rates of the extended SIFT matching algorithm were very low and first mark-recapture studies on a population of giraffes based on *Wild-ID* were promising. To date, *Wild-ID* has been successfully applied to other patterned animals such as salamanders [236] and wildebeests [237].

[4] `http://software.dartmouth.edu/Macintosh/Academic/Wild-ID_1.0.0.zip` Last visit: November 23^{rd}, 2013

Recently, Crall *et al.* presented a framework called *HotSpotter*[5] for individual identification of animals in [180]. Similar to *Wild-ID*, the system is not limited to a certain species and employs state-of-the-art algorithms for local keypoint detection and description. Although SIFT features are employed as well, a number of modifications were proposed which eventually lead to a more accurate and robust identification system. Moreover, the authors propose significantly improved strategies for matching two images (one-vs-one matching) as well as for matching a query image against the entire database (one-vs-many matching). The one-vs-one matching approach is comparable to the strategy applied in *Wild-ID*. However, instead of the brute-force matching paradigm used in [179], Crall *et al.* propose to build a forest of kd-trees [238], a data structure that organizes data points in k-dimensional space. This procedure not only speeds up the matching process but also improves its accuracy. For the one-vs-many strategy, the authors adopt a matching algorithm called Local Naive Bayes Nearest Neighbor (LNBNN), originally proposed by [239] for image classification. For details the interested reader is referred to [239, 180]. *HotSpotter* was successfully tested for a variety of species including zebras, jaguars, giraffes, and lionfish. Figure 3.11 shows a number of images that were correctly matched by *HotSpotter*.



(a) (b) (c) (d) (e)

Figure 3.11.: ***A variety of images of different species that were correctly matched by*** **HotSpotter.** (a) Giraffe, (b) Jaguar, (c) Lionfish, (d) Grevyzebra, and (e) Plain zebra. Image source: [180]

On average about 98% of the correct individuals were among the first five identities proposed by the system. Therefore, it consistently outperforms the *Wild-ID* system developed by Bolger *et al.* in [179].

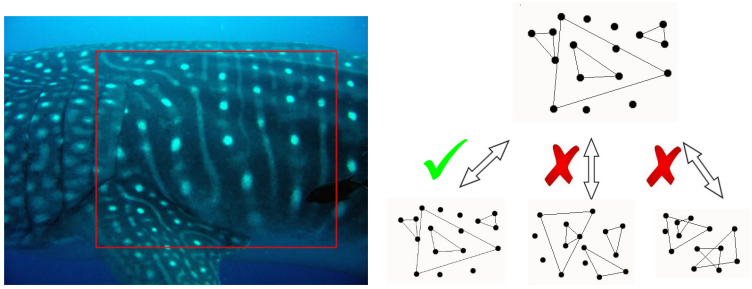
Another application of SIFT features was recently presented by Town *et al.* in [183]. The authors developed a framework called *MantaMatcher*[6] to identify manta rays by means of sparse pattern-based recognition methods. Unique natural spot patterns on the ventral surface area of mantas are utilized for identification. As in *Wild-ID* and *HotSpotter*, user interaction is required

[5] `https://github.com/Erotemic/hotspotter` Last visit: November 23rd, 2013
[6] `http://www.mantamatcher.org/` Last visit: November 23rd, 2013

to select a region of interest before keypoints and descriptors are extracted. A number of automatic image enhancement techniques are applied before feature extraction using a combination of median filtering and Contrast Limited Adaptive Histogram Equalization (CLAHE) [240] to remove noise and normalize lighting. Town *et al.* enhance the original SIFT matching approach by considering only features if the distance between the nearest neighbor and the second nearest neighbor is above a certain threshold and if keypoints were detected in a similar scale. Therefore, this procedure implicitly takes the size of the spot patterns into account. Town *et al.* compared their approach to other SIFT-like descriptors such as SURF [52] and ORB [230]. However, on a self-established dataset which consists out of 720 images of 265 manta-rays, SIFT in combination with the proposed pre-processing techniques and enhanced matching procedure outperforms SURF and ORB with a rank-1 recognition rate of approximately 51%. Despite the fact that all three systems perform well for a variety of patterned species, still human interaction is required to specify an image region which ideally contains only the animal of interest. However, as shown in Section 3.2.1 a variety of algorithms exist that are capable of reliably detecting the body of animals and even regions of interest on their bodies under real-world conditions.

While SIFT-like descriptors are solely based on gradient magnitudes and directions around certain points of interest, the information provided by the relative spatial positions of spot patterns on the body of animals is usually not taken into account. Therefore, researchers also adopted and developed sophisticated algorithms which explicitly exploit the spatial relationship between keypoints. One of the first implementations of a pattern-matching algorithm by means of spatial configurations of spot patterns was developed by Arzoumanian *et al.* in [241]. The authors adapted a computer vision algorithm originally developed by [242] for the astronomical community in order to compare stellar patterns of the night sky. Every combination of triplets of detected points within an image describes a triangle with particular scale and rotation invariant geometric properties. To compare two images of whale sharks, first all possible triangles within a query and reference image are formed. Subsequently, geometrically similar pairs of triangles are identified by comparing the extracted geometrical properties such as angles between vertices and relative side lengths. Repeating this procedure for every query-reference pair, images of individuals in the dataset can be ranked and eventually be reliably identified. Figure 3.12 illustrates the proposed matching procedure for an example image of a whale shark.

However, due to various image distortions such as noise, obscuration, reflected sunlight, and compression artifacts, detection of white spots on the surface of whale shark bodies requires a significant amount of manual interaction which according to [241] takes about ten minutes per image. However, it was shown that the proposed algorithm has a very low false positive rate while false classifications are mainly caused by large pose variations. The overall accuracy of the system was more than 90% on a long-term established dataset. Hence, the developed



(a) (b)

Figure 3.12.: An example image of a whale shark and the pattern matching approach proposed in [241]. (a) An image of a whale shark and a selected region of interest. The identification process is based on the spot coloration on the shark's body. (b) A stellar pattern matching algorithm based on the formation of triangles is exploited for individual identification. Geometrical properties of extracted triangles based on the spots of the test image are compared with triangle formations in the library. Only subsets of possible triangle configurations are shown. Figure adapted from [241]

algorithm is part of the web-based *ECOCEAN Whale Shark Photo-Identification Library*[7]. In 2007, Tienhoven *et al.* extended the ideas of [241] and proposed an interactive pattern-matching system for the identification of sharks [243]. The method uses an initial two-dimensional affine transformation based on three manually defined reference points to compare pairs of markings of two individuals in a commonly defined reference space. Spot pairs of aligned images are only defined as matched if the nearest alternative spot has at least twice the distance of the current match. The normalized sum of euclidean distances of matched spot pairs serves as distance metric in order to rank each shark image in the database with respect to the query. On a dataset gathered in a long-term study over eight years, the system has proven to be a useful identification aid achieving an accuracy of almost 92%, provided that more than three images per individual are present in the dataset. The proposed algorithm is part of a computer-assisted software solution for individual identification of patterned animals named *Interactive Individual Identification System (I^3S)*[8]. Both systems, the approach by [241] as well as the method developed by [243], have been successfully applied for non-invasive capture-mark-recapture studies conducted by Holmberg *et al.* in [244]. The authors used both systems to estimate the survival and recruitment rates of whale sharks in western Australia and demonstrated the direct scientific benefit of the proposed computer-assisted animal identification techniques. Furthermore, a thorough evaluation of I^3S [243] was conducted by Speed *et al.* in [245]. The authors not only determined the effect of different viewing angles but also how the number of spot pairs in matched images affects the system's performance. Although it was shown that the accuracy

[7] `http://photoid.whaleshark.org` Last visit: January 1^{st}, 2014
[8] `http://www.reijns.com/i3s/index.html` Last visit: January 1^{st}, 2014

significantly decreases as the viewing angle grows, the authors state that the software provides a reliable and freely available non-invasive solution for individual identification of wildlife. Moreover, Speed *et al.* claim that I^3S is not restricted to identify whale sharks but can be applied to a variety of spot patterned animals.

Inspired by ideas of [243], Anderson *et al.* developed a system to identify polar bears based on whisker spot patterns in [246]. The proposed algorithm exploits the individual specific arrangement of black spots originating from the base of whisker follicles. As done in [241] and [243], the user is required to manually select three reference points for affine image alignment. After cropping the desired image region, a number of state-of-the-art image enhancement techniques such as contrast enhancement, noise removal, and histogram equalization are applied. Subsequently, a binary image is obtained by adaptive thresholding. The Chamfer distance [247] is applied to compare two images which computes the euclidean distance between every black pixel of the query image to the nearest black pixel of the reference image. The median of all distances eventually defines the Chamfer distance and is used to rank the images in the dataset. Therefore, in contrast to the matching technique proposed by [243] which searches for the best matching spot pairs, Anderson *et al.* exploit all detected black spots within a query and a reference image. Hence, the approach by [246] might be more influenced by spurious pixels, i.e. pixels that were falsely detected as identification spots, than the technique used by Tienhoven *et al.* [243]. The proposed algorithm for identification of polar bears was evaluated on a dataset containing more than 200 individuals. However, only images that are suitable for identification according to various subjective quality criteria were selected for the experiments, resulting in a final testset of 57 individuals. On this high-quality dataset the system achieved reasonable results with a mean true positive rate of 80% at a false positive rate of 10%.

As already stated, one drawback of the spot matching systems proposed by [241], [243], and [246] is the strong dependence on user interaction such as image alignment and cropping, contrast enhancement, as well as noise reduction. To overcome these limitations, Albu *et al.* proposed a computer vision based system to identify spotted snakes with minimal user intervention in [248]. The approach efficiently exploits the unique spatial distribution of black spots underneath the head of the snakes for identification. The only required user intervention is the segmentation of the head to minimize the influence of the background. However, the authors claim that this manual pre-processing step will be eliminated in a future release due to a new image acquisition protocol that will help to gather higher-quality images which eventually leads to a full-automatic system. Once the snake's head is segmented, the five most prominent black spots located on its ventral side are detected using Otsu's automatic binarization technique [210]. A 3-dimensional feature vector containing geometrical characteristics of the spot distributions are subsequently extracted for every image and used for classification. Besides information

such as relative distance and orientation of the points to each other, one element of the obtained feature vector represents the area covered by a spot. Therefore, unlike the approaches by [241], [243], and [246] the approach by Albu *et al.* implicitly takes the size of the detected spots into account which might provide useful information for pattern matching, especially for markings with significant shape or size variation. A minimal distance classifier in combination with the k-means algorithm for localization of cluster centroids in the feature space is used to get the final decision about the individual. Unfortunately, the proposed algorithm was evaluated on a small dataset of only five individuals, achieving promising results of more than 80% accuracy. However, to achieve acceptable results on larger datasets, more sophisticated visual cues such as the shape of the spots for instance have to be taken into account.

One of the most sophisticated and completely autonomous systems for visual identification of animals was proposed in [249, 130]. Burghardt *et al.* apply algorithms for animal detection, 3D model posing, and biometric identification to recognize African Penguins by means of unique deformable coat patterns on their chests. Unlike most other automatic approaches for animal identification, the developed system can not only be applied to recognize individuals in photographs but rather in videos filmed in widely unconstrained, natural habitats. To obtain a coarse estimate of the object location and size of the penguin, Burghardt *et al.* use Haar-like descriptors and AdaBoost in combination with a tracking technique originally published in [105, 106] (see section 3.2.1). Based on the initial detection results, the pose of the animal is estimated by fitting a deformable 3D model to the surface of the detected animal. Again, Haar-like features are utilized to robustly model the penguin's local appearance and detect keypoints and surrounding patches on the animal's body surface. The spatial coherence of patches is modeled by a probabilistic spatial prediction model termed *Feature Prediction Tree* [249, 130]. Starting from a root node, a particular keypoint on the center of the top chest of penguins which can be reliably detected in all poses of interest, the Feature Prediction Tree models possible pathways to other pre-defined keypoints in a flexible manner. Following this procedure, the correspondences between model and image features can eventually be reliably derived to pose a 3D model into the scene. Although this technique is prone to occlusions, it performs well in situations where the chest of penguins is abundantly visible which is a prerequisite for accurate identification. Once the 3D model is fitted, the spot patterns on the penguin's chest can then be back-projected into 2D space which generates a normalized aligned texture map used for further processing. Spot locations are subsequently found by detecting phase singularities as proposed in [130] (see Section 3.2.1). Since the spatial distribution of the extracted markings still suffers from non-linear local deformation of the skin, the authors furthermore propose a robust matching strategy based on *Shape Contexts* introduced by Belongie *et al.* in [250]. The method builds polar histograms of feature positions with increasing bin size further away from

a reference point. Burghardt *et al.* extend these ideas by replacing the actual feature location with feature distributions modeled by a Gaussian to describe positional uncertainties relative to the reference marking. In the implementation of [249] and [130] every detected landmark point on the chest of a penguin takes the role of the reference spot once which eventually leads to the construction of multiple polar histograms called *Distribution Context*. For the actual matching process of two Distribution Contexts, the Hungarian method [251] is utilized to solve the assignment problem and the sum of associated matching costs is finally used as distance metric to compare two patterns. A proof-of-concept of the developed *African Penguin Recognition System (APR)* was presented by Sherley *et al.* in [130, 252]. The system was thoroughly evaluated at a native colony of African Penguins on Robben Island, South Africa. Although only 193 individuals out of 1453 birds that passed the camera during a selected time interval produced frontal detections that were suitable for subsequent identification, a matching sensitivity of up to 96.7% at a low false acceptance rate of 10^{-2}% was achieved. To proof the broad applicability of the proposed system, a preliminary performance study was conducted on plain zebras in [130]. On a dataset of 200 images of 47 individuals the framework also achieved satisfying results with a genuine acceptance rate of 92% at 0.1% false acceptance rate.

As shown by the above mentioned approaches, the individual visual uniqueness of natural coat patterns can provide robust features for computer-assisted identification of animals for a variety of different species. However, such approaches are often infeasible for the identification of animals without obvious visual markings because individually unique patterns are either not existent or cannot be exploited due to limited image or video resolution. Moreover, some studies have reported drastic changes of natural markings on individuals of certain species during an animal's natural growth or evolution, indicating that animals can change their appearance significantly [253, 254, 255, 256]. Thus, computer vision algorithms solely based on coat patterns might be ineffective if coat patterns of animals change significantly over time. This is also true for contour-based methods reviewed in Section 3.2.2 since mostly artificially obtained markings such as scars, wounds, and scratches are utilized for identification which might transform or completely heal over time [257, 258, 255, 259]. Furthermore, some of the introduced algorithms for pattern-based identification (e.g. [179, 184, 180]) can only be applied if the animal is positioned in a left or right profile view, respectively; i.e. if descriptors were extracted from the left side of an animal's body during training, the individual can only be re-identified if filmed in a left side posture and vice versa. This is due to the fact that the left and right sides of bilateral animals are in general not symmetric [260]. To overcome these issues, a small number of attempts have been made to adapt automated face recognition techniques, originally developed to recognize humans, in order to identify animals.

Methods based on Face Recognition

Whereas automatic face recognition for humans has been a major research topic for decades, only a few publications can be found that extend and adapt these techniques to identify animals.

In 2005, Kim *et al.* proposed a system for cattle identification in livestock farms to overcome the limitations of invasive methods such as RFID (Radio Frequency Identification) implants in [261]. Previous non-invasive approaches to identify cattle by means of image processing focused on black-and-white patterns or muzzle prints [186, 187, 188]. However some sub-species such as the Japanese Black Cattle for instance do not carry such unique markings. Therefore, the authors propose to utilize a face recognition technique based on gray-level intensities and neural networks to discriminate between individuals. Kim *et al.* gathered one face image of 12 cows each at feeding time, i.e. under relatively controlled conditions, to evaluate the proposed system. To test the systems robustness to various extrinsic factors, the authors manually modified the obtained images by adding noise, rotations, distortions, and lighting changes. Although faces could be re-identified by the system under moderate image modifications, a thorough evaluation is missing since a valid dataset of different individuals photographed under various natural conditions needs to be established. Moreover, the authors come to the conclusion that the proposed algorithm is not suitable to recognize moving cows in real-life scenarios but could instead be used to identify stationary cattle at the feeding area for instance.

A second application of face recognition algorithms for animal identification was published by Corkery *et al.* in [262]. They presented a preliminary study of a holistic face recognition approach based on Independent Component Analysis (ICA) [263] to identify sheep. To compensate adjustments in scale, rotation, and lighting, a number of pre-processing steps were applied to normalize the facial images before the actual recognition. Based on manually annotated eye coordinates, each face image is first rotated into an upright position and subsequently cropped. Furthermore, images are smoothed by convolution with a Gaussian kernel to remove noise and a gray-level histogram equalization is applied to compensate illumination variation. After vectorizing the normalized gray-scale face images, Principal Component Analysis (PCA) [61] is exploited to transform the high-dimensional feature vectors into a smaller subspace. Based on the work of [263], ICA is subsequently used for face representation. In ICA, a method for blind source separation, the goal is to decompose an observed signal into additive statistically independent subcomponents which are used as basis images to represent faces. A simple nearest neighbor classifier with cosine distance is used for the final classification. For evaluation purposes, multiple face images of 50 individuals were gathered in a controlled environment in order to minimize the influence of undesired lighting conditions. The algorithm obtained recognition rates up to 95%. Although the proposed system performed well, images were taken in a rather controlled setting, suggesting that the accuracy will significantly decrease under real-world con-

ditions. Moreover, human interaction is necessary to detect facial markers to align the facial images.

Based on the assumption that humans and our closest relatives, the great apes, share similar properties of the face, this thesis explores the feasibility of face recognition methods for the identification of individuals as well. However, unlike the approaches of [261] and [262] a completely automatic non-invasive unified framework for detection and identification of chimpanzees and gorillas is presented.

It was shown in the previous section that a variety of different algorithms were developed in the recent past to identify animal individuals by means of automated computer vision methods. Table 3.2 gives an overview of sate-of-the-art methods for non-invasive animal identification in image and video footage.

Category	Reference	Species	Description	Notes
Contour-Based Methods	Hillman *et al.* [195]	Dolphins	(1) Smooth approximation of fin's edge based on polynomial function (2) Extraction of low-level descriptors (normalized peak position, width & height of peaks) (3) Affine-invariant Fourier descriptors	(1) Manual annotation of ROI and the tip of the dorsal fin (2) Thorough evaluation is missing
	Kreho *et al.* [197]	Dolphins	(1) LoG edge detection (2) Curve fitting by Active Contour Model (3) Extraction of *Dorsal Ratio*	(1) Manual selection of start and end point fin's edge (2) Significantly reduced processing time compared to manual analysis
	Araabi *et al.* [200]	Dolphins	(1) LoG edge detection (2) Active Contour Model for curve fitting (3) Extraction of low-level attributes of curvature after zero crossing (width, height, area of positive and negative parts)	(1) Extraction of more attributes than previous methods (2) Outperforms previous approaches (3) Semi-automatic approach ($\sim$ 5 minutes processing time per image) (4) Not robust against affine distortions

Category	Reference	Species	Description	Notes
	Gope *et al.* [201]	Dolphins Sea Lions Whales	(1) Interactive edge detection (2) Fitting spline curve for approximate curvature extraction (3) Affine transformation of extracted curvatures (4) Area of mismatch between two transformed curves serves as distance measure	(1) System robust to affine transformations (2) User interaction required for edge detection (3) Proposed system not compared to previous approaches
	Ardovini *et al.* [203]	Elephants	(1) Identification based on characteristic nicks in ears (2) Matching of non-connected curve sets (3) Canny-edge detection algorithm (4) Two-stage matching strategy: (a) Global matching based on keypoints, (b) Shape comparison of globally matched curve sets	(1) Manual annotation of reference points as well as start and end points of nicks (2) Robust to partial occlusion

Category	Reference	Species	Description	Notes
Dense Pattern-Based Methods	Hiby *et al.* [181, 182]	Seals Sharks Zebras Lynx	(1) Based on pelage markings on fur and skin (2) Fitting 3D model of animal body for alignment (3) Gray-scale intensities on grid-cells of fitted model as descriptors (4) Matching of two images based on correlation-coefficient between corresponding elements (5) Method is extended in [182] by combining two complementary features	(1) Manual annotation of keypoints necessary for initial 3D model placement (2) User interaction required to refine the model fitting (3) Program available for download at `http://www.conservationresearch.co.uk/`
	Kehtarnavaz *et al.* [209]	Whales	(1) Black and white patterns on flukes for identification (2) Edge detection for fluke segmentation (3) Otsu's method for adaptive thresholding and binarization (4) Affine Moment Invariants as descriptors (5) Nearest Neighbor Classification	(1) Semi-automatic fluke segmentation (2) No alignment necessary due to invariant features (3) Outperforms contour-based method by Araabi *et al.* [200]
	Gope *et al.* [212]	Whales	(1) Extension of approach by [209] (2) Replacement of Affine Moment Invariants with Zernike Moment Invariants (3) LDA for features space transformation	(1) Semi-automatic approach for fluke segmentation (2) Approach outperforms system proposed by [209]

Category	Reference	Species	Description	Notes
	Ranguelova *et al.* [214, 215]	Whales	(1) Fitting affine invariant grid to segmented fluke (2) Relative contribution of black and white pixels in each grid cell serve as features (3) Nearest Neighbor Classification (4) Extension in [215] by fitting ellipse to every detected blob and take extrema points of ellipse as additional features	(1) Grid is fitted based on manually selected keypoints (2) Method is shown to be robust to various image qualities
	Krijger *et al.* [216, 217]	Zebras	(1) Preprocessing: deinterlacing, smoothing, blockwise binarization (2) Sequential thinning for skeletonization (3) Extraction of curvature, slope, and intersection points as features	(1) Manual selection of ROI based on 6 keypoints (2) Method not robust against affine transformation in 3D space
	Lahiri *et al.* [184]	Zebras	(1) Development of open-source framework for zebra identification called *StripeSpotter* (2) Extraction of *StripeCodes* based on coat patterns	(1) Manual selection of ROI necessary (2) Extracted *StripeCodes* robust to scale, exposure, and partial occlusion (3) Source code available at `http://code.google.com/p/stripespotter/`

Category	Reference	Species	Description	Notes
	Ravela *et al.* [220] Gample *et al.* [221] Duyck *et al.* [185]	Salamanders	(1) Deformable body has to be straightened for texture alignment (2) Multi-scale Gaussian derivative filters for feature extraction (3) Shape and orientation histograms for compact feature representation (4) Normalized cross-covariance for histogram matching	(1) Image acquisition in custom enclosure for constant background and lighting (2) Manual annotation of keypoints for ROI selection and alignment (3) Creation of collaborative community framework *MIT SLOOP* and inclusion of more species [225, 185]
	Hoque *et al.* [226]	Newts	(1) Identification based on dark spots on the belly of newts (2) Correlation coefficients on vectorized gray-scale intensities for matching	(1) Manual interaction necessary for texture alignment (2) Gray-scale information alone is often not robust enough against image distortions
	Azhar *et al.* [227]	Newts	(1) Extension of approach by [226] (2) Multi-radii LBP histograms as descriptors (3) Chi-squared distance for matching	(1) Manual user interaction required for texture alignment (2) Outperforms method proposed by [226]

Category	Reference	Species	Description	Notes
	Cannavò *et al.* [228]	Frogs	(1) Based on pigmentation patterns on dorsal side of the body (2) Otsu's thresholding method and morphological operations for body detection (3) Second order statistics defines orientation of detected body (4) Extraction of texture features such as granulometry [229] (5) Nearest neighbor classification	(1) Images gathered in controlled environment with constant background and lighting (2) Completely automatic identification system (3) Artificially modified images used for system evaluation
Sparse Pattern-Based Methods	Pauwels *et al.* [234] de Zeeuw *et al.* [129]	Turtles	(1) Identification based on *pink spot* on turtle's head (2) Image comparison based on SIFT matching (3) Extension in [129] by cross-correlation of gray-scale values	(1) Manual selection of ROI necessary
	Bolger *et al.* [179]	Multiple species (giraffes, salamanders, wildebeests)	(1) SIFT feature detection and description (2) RANSAC algorithm [235] to find geometrically self-consistent subset of keypoints	(1) Manual selection of ROI suggested to minimize influence of background (2) Open source implementation of *Wild-ID* can be found at `http://software.dartmouth.edu/Macintosh/Academic/Wild-ID_1.0.0.zip`

Category	Reference	Species	Description	Notes
	Crall *et al.* [180]	Multiple species (zebras, jaguars, giraffes, lionfish)	(1) Modifications of standard SIFT detection, description, and matching (2) Forest of kd-trees for efficient matching (3) Local Naive Bayes Nearest Neighbor (LNBNN) algorithm for keypoint matching	(1) Manual selection of ROI (2) System outperforms *Wild-ID* [179] (3) Open source implementation of *HotSpotter* can be found at `https://github.com/Erotemic/hotspotter`
	Town *et al.* [183]	Manta Rays	(1) Identification based on spot patterns on ventral surface of animals (2) Median filtering and CLAHE for image enhancement (3) Extension of basic SIFT matching by taking scale and distances between keypoints into account	(1) Manual selection of ROI (2) Enhanced SIFT matching approach outperforms other local keypoint descriptors (3) Program available at `http://www.mantamatcher.org/`
	Arzoumanian *et al.* [241]	Whale Sharks	(1) Algorithm for stellar pattern matching used for matching spot patterns (2) Based on triangle matching	(1) Manual selection of spots used for matching (2) System is part of ECOCEAN whale shark photo-ID library `http://photoid.whaleshark.org`
	Tienhoven *et al.* [243]	Whale Sharks	(1) Extension of algorithm proposed by [241] (2) Initial affine transformation for alignment (3) Spot matching based on cartesian coordinates of detected keypoints (4) Sum of normalized Euclidean distances between images serves as similarity measure	(1) Manual selection of three reference points for alignment (2) Algorithm is part of *Interactive Individual Identification System* I^3S `http://www.reijns.com/i3s/index.html`

Category	Reference	Species	Description	Notes
	Anderson *et al.* [246]	Polar Bears	(1) Identification based on whisker spot patterns (2) Image alignment using affine transformation (3) Image enhancement to compensate noise and lighting changes (4) Adaptive thresholding for binarization (5) Chamfer-distance [247] for global pattern comparison	(1) Manual selection of three reference points for alignment (2) Method might be prone to errors caused by spurious pixels
	Albu *et al.* [248]	Snakes	(1) Method exploits spatial distribution of black spots on bottom-side of snake heads (2) Automatic location of 5 most prominent key-points (3) Geometrical characteristics of spot distributions used as features (4) Minimal distance classifier in combination with k-means clustering used for matching	(1) Manual segmentation of snake heads (2) New image acquisition protocol might lead to full-automatic system in the future (3) Evaluation only on small dataset of 5 individuals

Category	Reference	Species	Description	Notes
	Burghardt *et al.* [249, 130]	African Penguins	(1) Completely automatic system (2) Deformable coat patterns on penguin chests is used for discrimination (3) Penguin detection using Haar-like features and AdaBoost (4) Object tracking in video-sequences [105, 106] (5) Pose estimation and 3D model fitting by means of Feature Prediction Trees (FPT) (6) Detection of spots based on generic phase curl localization [131] (7) Distribution Contexts based on polar histograms of distributions of spot locations as descriptors (8) Hungarian method and sum of associated matching costs used to compare chest patterns	(1) No user interaction required (2) Real-world prototype and proof of concept evaluated in [252] (3) Preliminary study on zebra identification (4) High accuracy in real-world environments at a low false acceptance rate (5) Only a fraction of animals produced detections which were suitable for subsequent identification
Face Recognition	Kim *et al.* [261]	Japanese Black Cattle	Face Recognition based on gray-scale information and Artificial Neural Networks (ANN)	(1) Manually cropping head region (2) System evaluated on artificial dataset (3) Proposed method not suitable to identify cattle in real-world environments

Category	Reference	Species	Description	Notes
	Corkery *et al.* [262]	Sheep	(1) Preprocessing for image enhancement (2) Gray-scale intensities and PCA for feature extraction and feature space transformation (3) Independent Component Analysis (ICA) and Nearest Neighbor Classification based on cosine distance	(1) Manual selection of eye coordinates for face alignment and cropping (2) Dataset gathered in restricted environment

Table 3.2.: ***Overview of state-of-the-art algorithms for non-invasive animal identification in images and videos.***

Automatic Species Classification

Besides individual identification of animals in audio-visual footage, the development of automatic routine procedures to reliably classify the species of detected animals arose interest of researchers and computer scientists within recent years. Taxonomic classification of animals is a prerequisite for many biological questions such as biodiversity conservation and natural resource management [264]. However, to date only few computer vision algorithms have been proposed in the literature to classify the species based on their morphological traits. The development of computer vision algorithms to automatically identify the species of an animal is not a trivial task because

1. Individuals of a given species might differ drastically in their morphology due to phenotypic variations caused by age or environmental conditions and

2. Different species might have similar morphological traits because many taxonomic groups often comprise thousands of species [265].

Enormous amount of progress has already been made in the field of automatic insect classification. Three of the most promising systems which are commonly used by experts to reduce the burden of manual classification of specimens are *ABIS* [266], *SPIDA-web* [267], and *DAISY* [268]. Within this thesis, a coarse outline of these systems as well as more recent approaches is given. For a state-of-the-art overview of the above mentioned animal classification frameworks the interested reader is referred to [265, 264, 128, 269].

The *Automated Insect Identification System (ABIS)* was one of the first sophisticated approaches for automatic taxonomic categorization of bees based on the venation of their wings [266, 270, 271]. Each bee is manually positioned under a microscope with background illumination in standard pose. For classification, the system follows a hierarchical approach by first determining a set of key wing cells, the area between veins, based on line and intersection detections. A set of low-level descriptors is subsequently extracted for initial classification in order to select a certain pre-defined deformable venation template saved in an external database. Once the template is fitted to the wing image, remaining cells can be reliably detected and the previously extracted feature vector is extended with a number of features obtained from the intensity values within a sampling window [270]. A Support Vector Machine (SVM) and Kernel Discriminant Analysis (KDA) are finally used for classification. The system is known to perform well even for bee species that are known to be hard to classify even for human experts. Although the features used to classify bee species are known to perform well, they make the system very specialized to a certain kind of insect. Another system commonly used by experts to distinguish between different spider species is called *Species Identification, Automated and web-accessible (SPIDA-web)*[9], introduced by Russel and Do in [267, 272]. The proposed algorithm utilizes Artificial Neural Networks (ANN) to classify spiders based on their external genitalia. Direct user interaction is required to annotate the region of interest within an image gathered under a microscope with constant background lighting. The wavelet coefficients of Gabor filters are used as input for multiple back-propagation neural networks trained for each species separately. Preliminary results for female spiders presented in [267] indicate that *SPIDA-web* is capable to achieve accuracies up to 95%. Although it was claimed by the authors that the system was created as a generalized classification system it was to date only tested on spiders. To overcome these limitations, O'Neill *et al.* proposed a generic identification system based on pattern recognition called *Digital Automated Identification System (DAISY)*[10] in [268, 273]. While the first version of *DAISY* [274] exploited the Eigenfaces approach originally developed for human face recognition [61], the core classification algorithm of the recent version of *DAISY* is based on a neural network approach called Plastic Self Organizing Maps (PSOM) [273]. The system has been successfully applied to a variety different insect species such as bumblebees, moths, wasps, midges, butterflies, larvae, and spiders. Although the results achieved by *DAISY* are comparable to the performances of specialized systems such as *ABIS* and *SPIDA-web*, user interaction is still required to manipulate the specimen, capture the image, and segment the region of interest which ultimately hampers the throughput of these systems for large datasets.

[9] `http://research.amnh.org/iz/spida/common/index.htm` Last visit: January 17^th, 2014
[10] `http://www.tumblingdice.co.uk/daisy` Last visit: January 17^{th}, 2014

In the recent past, a considerable amount of literature has been published on insect classification to overcome the limitations of *ABIS*, *SPIDA-web*, and *DAISY*. An autonomous system for bee classification named *DrawWing*[11] was proposed by Tofilsky in [275]. Unlike *ABIS*, Tofilsky's approach is not only based on standard morphometry of wing venation but also on characteristic landmark points, so called geometric morphometrics. Since points of interest are automatically located by the software based on vein junctions, no user interaction is required to align wing images. However, the wings should be detached from insects bodies before image capturing to achieve good results, which not only requires a significant amount of user interaction to prepare the specimen but more importantly harms the animal before classification. In [128], Larios *et al.* proposed a combined hardware-software system for automatic taxonomic insect classification using histograms of local appearance features. To automatically categorize stonefly larvae, the authors developed a mechanical system for photographing the specimens under a microscope and provided a software tool for region detection, feature extraction, and classification. Larios *et al.* propose a Bag-of-Words (BoW) approach [276] where in a first step three different detectors are used to locate regions of interest. Each detected region is subsequently represented as SIFT features [51] and k-means clustering over all descriptor vectors is applied to build the keyword dictionaries. Codewords are then defined as the centers of the learned clusters. Thus, each detected patch of the larvae is mapped to a certain codeword through the clustering process. Each input image can therefore be represented by the histogram of codewords. The combination of three different detection algorithms boosts the performance of the algorithm significantly and outperforms systems solely based on only one of the used detectors, achieving accuracies of up to 82% for a four-class problem. A similar approach named *BugID* was used by Lytle *et al.* [277] for automatic classification of benthic invertebrate samples. Similar to the approach by [128], Lytle *et al.* apply three different region detectors to localize object instances at different scales and shapes. Each detected region is represented by SIFT features [51] and subsequently compared to samples in the database using a random forest matching approach [278]. However, while previous approaches for insect classification treat the problem in a closed-set fashion, Lytle *et al.* propose to utilize an open-set classification scheme. While in closed-set classification it is assumed that all possible classes are known to the system, an open-set system first has to decide if a probe is known to the system before it is actually assigned to a certain class of the training set (see Section 5.3 for details). Such a procedure helps biologists to quickly identify novel species not yet present in the database which, according to [277], is the common scenario for most field-collected samples. The experiments on a dataset of 9 different species shows that 94.5% of correctly accepted samples was classified correctly while most unknown species were correctly rejected by the system. In 2012, Utasi proposed a local appearance feature based ap-

[11] `http://drawwing.org/` Last visit: January 17^{th}, 2014

proach for automatic categorization of tarantulas in [279]. While *SPIDA* focuses on the external genitalia of spiders to classify the species, Utasi *et al.* followed a more general approach by using local color descriptors known as colorSIFT [280], a variant of SIFT applied to different color channels. After feature extraction, the Bag-of-Words (BoW) model [276], a histogram representation of visual features, is adopted to obtain a more compact and meaningful representation. In [279], the authors compare different state-of-the-art classification methods such as Naïve Bayes Classification, linear Support Vector Machines (SVM), and Supervised Latent Dirichlet Allocation (sLDA) [281], were the latter performed best with an accuracy up to 77% on a dataset of 7 different species.

Although the majority of proposed algorithms for species classification are specialized to categorize insects, there has been an increasing amount of literature on classification of larger animals such as fish, reptiles or mammals in recent years. In 2010, Rodrigues *et al.* presented an approach for automatic classification of fish species in [282]. For feature extraction, the authors propose to apply PCA on vectorized color channels in the YUV color space to encode both brightness and color information. Next, unsupervised clustering techniques based on two immunological algorithms, Artificial Immune Network (aiNet) [283] and Adaptive Radius Immune Network (ARIN) [284], are utilized in order to find natural groupings of features extracted from different individuals of different species. By this means, the resulting clusters refer to features gathered from individuals of the same species. Finally, a Nearest Neighbor Classifier based on the distance between a new feature vector and the obtained cluster centroids is used for classification. The proposed algorithm is evaluated on a dataset of 4 different fish species and compared with a SIFT matching approach. Although the system outperformed SIFT matching and achieved an overall accuracy of 92%, several images of only one individual per species was present in the dataset which makes it hard to estimate the generalization capability of the system. Furthermore, the dataset was gathered in a controlled environment with constant background to eliminate variations resulting from background clutter and challenging lighting conditions which are often present in a real-world environment. Spampinato *et al.* proposed a complete system for fish detection, tracking, and classification operating in natural underwater environments in [285]. For detection and tracking, the authors used an approach developed in their earlier work [286] (see section 3.2.1). Based on the regions of interest obtained from detection and tracking, the authors subsequently extract a number of texture and shape descriptors. The texture of the object is described by statistical moments of its gray-scale histogram, Gabor wavelets, and various properties of the gray-level co-occurrence matrix. Shape information on the other hand is characterized by histograms of Fourier descriptors of the object boundary obtained from the Curvature Scale Space (CSS) image proposed in [287]. PCA and Discriminant Analysis are finally applied for feature space transformation and classification, respectively.

An average classification accuracy of 92% was achieved on a dataset of 10 different fish species gathered under real-world underwater conditions which shows the applicability of the proposed algorithm.

Also the classification of mammal species has drawn the interest of a number of computer vision experts. Kouda *et al.* [288] for instance proposed a system to reliably differentiate between raccoons and raccoon dogs which look very similar in shape and appearance. The authors developed an intelligent camera trap based on face detection and recognition techniques for population monitoring for these two species. Face detection is based on Histogram of Oriented Gradients (HOG) [104] in combination with a Support Vector Machine (SVM) for classification. However, according to Kouda *et al.* , a reliable differentiation between raccoons and raccoon dogs solely based on the detection confidences is not feasible. Therefore, the authors trained a second SVM for species classification. Feature vectors were obtained from the coefficients of the Discrete Cosine Transform (DCT) of the input image. Furthermore, a feature selection algorithm was applied before classification in order to pick the coefficients which are best suited for discrimination. Another study by Wilber *et al.* introduced techniques for animal detection and classification that can help biologist to study squirrels and tortoises in the Mojave Desert using mobile devices [289]. For animal localization, the authors apply the keypoint detection algorithm used in SIFT [51] to extract a number of sparse keypoints. Around each detected point of interest a LBP [50] based descriptor is extracted and a 1-class SVM is applied to distinguish target objects (squirrels and tortoises) from objects that are not of interest. For species classification, Gabor features are extracted from the automatically obtained regions of interest and a multi-class SVM is used to differentiate between three different squirrel species. Results on a self-established dataset shows the effectiveness of the proposed algorithm with an average recognition rate of around 78%.

Afkham *et al.* on the other hand proposed to use joint visual texture information of detected animals and their background for animal classification [290]. Therefore, an additional segmentation algorithm to extract the animal from the background is obsolete. The authors apply a method adopted from visual object categorization based on a visual word dictionary generated from Markov Random Field (MRF) descriptors [291]. Furthermore, Afkham *et al.* propose to apply a joint probabilistic model in order to obtain more discriminative features. The main idea of applying joint probabilities is to capture the likelihood that different visual words appear in the neighborhood of each other which implicitly encodes the information about context and the background surrounding the object. The proposed approach achieved promising results on a publicly available dataset of $1,239$ images of 13 animal species. A particularly interesting and sophisticated approach for automatic animal categorization of wildlife pictures captured by remote camera traps was developed by Yu *et al.* in [292]. The system exploits the sparse

coding spatial pyramid matching (ScSPM) paradigm proposed in [293], a method for general object categorization. The algorithm first extracts dense SIFT features and LBP descriptors on a manually cropped image region. The weighted sparse coding scheme for dictionary learning, originally proposed by Wang *et al.* in [294], is utilized to generate a compact dictionary that can sparsely represent the incoming descriptors with minimum error. Subsequently, spatial pyramid matching, an extension of the Bag-of-Words (BoW) approach, is used to model the spatial layout of local image features at multiple scales. A linear SVM is finally used for classification. On a challenging self-established dataset of 7,000 images of 18 species gathered under natural conditions from autonomous camera traps, remarkable results could be achieved by the system with an average recognition rate of 82%. However, although a variety of algorithms for automatic detection of animals in video footage exist (see section 3.2.1), manual segmentation of the animal is still required for the proposed framework.

3.3. Face Recognition

As stated earlier, the proposed Primate Recognition Framework (PRF) applies face recognition algorithms to identify great ape individuals. Hence, a comprehensive overview of state-of-the-art face recognition algorithms is given in this section. The proposed PRF is benchmarked against traditional and more recently proposed face recognition algorithms in Chapter 5 to show the superiority of the developed algorithm for the identification of great apes. Thus, particular emphasis is placed on face recognition methods that are compared with the proposed framework, while others are only described briefly.

Automatic face recognition is one of the most fascinating and challenging problems in the field of computer vision and image understanding. It has therefore received significant attention by many researchers and computer scientists throughout the past three decades. Major improvements were made since the first attempts in the early 1990s. Thus, automatic face recognition has become one of the most successful applications of image analysis. The reasons for this trend are twofold: First, feasible technologies for automatic face recognition in images or videos have a wide range of many commercial and law enforcement applications such as advertising, market research, and surveillance. Secondly, after more than 30 years of research, significant progress has been made in the field of automatic face recognition and biometric identification and thus algorithms for robust and accurate identification are nowadays commercially available. Table 3.3 gives an overview of some commercial face recognition systems.

Commercial Product	Company	Website
FaceVACS	Cognitec	http://www.cognitec.com/
Face++	Megvii Inc.	http://www.faceplusplus.com/
SEKUFace	EUROTECH	http://www.eurotech.com/en/products/SekuFACE
FaceExaminer	MorphoTrust	http://www.morphotrust.com/Technology/FaceRecognition.aspx
IWS Biometric Engine	ImageWare Systems	http://www.iwsinc.com/
BioID	BioID AG	https://www.bioid.com/
Visual Casino 6	Biometrica	http://biometrica.com/
MFlow Journey	Human Recognition Systems	http://www.hrsid.com/company/technology/face-recognition
Picasa	Google	http://picasa.google.de/intl/de/
IPhoto	Apple Inc.	http://www.apple.com/mac/iphoto/
ReKognition	Orbeus	http://www.rekognition.com/

Table 3.3.: ***Available commercial face recognition systems for surveillance and entertainment.*** Note that some of the links might have changed. Last visit of all websites: April 24^{th}, 2014.

Face recognition applications itself are mainly used for three tasks:

1. Verification (one-to-one matching): The face recognition system has to determine if the person in an image is who he/she claims to be.

2. Identification (one-to-many matching): Out of a limited set of classes, the face recognition system has to determine the identity of the person in the image.

3. Watch-List: In an open-set classification scheme, the face recognition system first has to decide if a person is known or unknown, i.e. is part of the training set. If not, it has to reject the person as impostor. If, however, the person is known to the system it has to determine his/her identity.

Automatic visual facial analysis and face recognition in particular is one of the most challenging tasks in object recognition. First, although rigid object detectors are commonly used to localize faces in images and videos the human face generally is a highly deformable object. Different facial expressions for instance can vary the visual appearance of a face significantly. Moreover, numerous other extrinsic and intrinsic factors might cause the appearance of a face to vary [295]. Intrinsic factors are purely influenced by the physical nature of the face such as age and different facial expressions. Extrinsic factors on the other hand arise from outside of the

individual. These factors include illumination, pose, occlusion, scale, and imaging parameters such as resolution, focus or noise among others. A facial recognition system should be robust against those kinds of image variation and a dataset for thorough evaluation should include such factors. However, face recognition algorithms are often evaluated on official benchmark datasets gathered under laboratory conditions where they usually perform quite well. Once such techniques are applied in real-world environments, often a significant decrease in performance can be observed. However, the recently published Labeled Faces in the Wild (LFW) dataset [296][12] provides a challenging benchmark dataset to test face recognition approaches under unconstrained conditions. An important part of face recognition research is the thorough evaluation and benchmarking of developed algorithms including a detailed reporting of the experimental setup. The Face Recognition Vendor Test (FRVT)[13] for instance provides independent evaluations of commercial and academic face recognition algorithms under challenging and realistic conditions using standard performance measures. This not only helps the face recognition community to identify future research directions but is also an opportunity for researchers to easily evaluate reported results and benchmark their systems against state-of-the-art algorithms as results become independently reproducible.

Machine vision for automatic face recognition has not only attracted interest of computer scientists but also of researchers from diverse scientific disciplines such as image processing, pattern recognition, computer vision, machine learning, computer graphics, and psychology. Thus and because of the vast and diverse amount of literature published in the field of automatic face recognition, it is difficult to find a generic taxonomy of face recognition algorithms. Moreover, a complete literature survey of existing face recognition techniques is out of the scope of this thesis. For a more complete and detailed overview of automatic state-of-the-art machine vision algorithms for human identification by means of their facial appearance the interested reader is referred to [297, 298, 299, 300, 301, 302].

According to [300] face recognition techniques can broadly be divided into three main categories based on their image acquisition protocol:

1. Algorithms that utilize data from multiple sensors, e.g. stereo cameras, infrared cameras or 3D sensors.

2. Methods that operate on intensity images obtained from a single camera.

3. Techniques that perform face recognition in videos.

[12] `http://vis-www.cs.umass.edu/lfw/` Last visit: April 22nd, 2014
[13] `http://www.nist.gov/itl/iad/ig/frvt-home.cfm` Last visit: April 15th, 2014

This review concentrates on the second category which is by far the most applicable one in real-world non-intrusive situations. Furthermore, for face recognition in video, often the techniques developed for still images are applied on a few selected frames after face detection and tracking [298]. However, recently approaches were proposed which incorporate multiple modalities in order to increase the performance of a face recognition system. Steffens *et al.* [303] for instance utilize stereo information to increase the robustness of the system while [304] exploits visual and audio information to build a multi-modal identification system. Another approach by Sivic and Everingham [305, 306] investigates the problem of automatically labeling characters in TV or movie material. Frontal faces are detected in every frame of the video sequence and are subsequently tracked through the shot. Each track is then represented by a set of feature vectors of local keypoint descriptors calculated around certain facial interest points. To associate each face track with a character name, a transcript of the movie is aligned with its subtitles by dynamic time warping and utilized as additional cue to support the results obtained by face recognition. A video-based face recognition which is solely based on image processing was presented by Ekenel and Stallkamp [307, 308]. Once faces are detected and tracked through the video sequence they are classified by a local appearance-based approach which is based on the Discrete Cosine Transform (DCT) applied on non-overlapping blocks of a face image. Classification is performed in every frame and results are combined by a frame weighting technique which is based on the classification scores. The system was evaluated on a database of 41 subjects and reached high accuracies even under difficult illumination conditions. However, the system was presented as a door monitoring system which limits the amount of possible face poses significantly, i.e. faces in a video sequence were mostly full-frontal which makes the recognition task relatively simple compared to an application scenario presented in this thesis.

Early approaches for image-based face recognition developed in the early and mid-1970s measured facial attributes such as the distances between certain facial landmark points and used them as unique features for identification [309, 310]. However, a precise automatic localization of facial landmarks is hard to achieve in practical applications. Moreover, simple distance measures are usually not robust enough against the intrinsic and extrinsic factors explained above. However, with the progress of statistical and machine learning techniques in the early 1990s, new interest in automatic face recognition approaches arose. Zhao *et al.* devide face recognition methods for intensity images into three main categories [299]: *Holistic Appearance-based Approaches*, *Local Keypoint Methods*, and *Hybrid Techniques*.

Holistic Appearance-based Approaches

Holistic approaches are one of the most successful and very well studied techniques for human face recognition. These methods use the whole face image as input to the recognition system. Many holistic techniques even use the raw gray scale pixel intensities as features. These methods are often based on subspace methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) or even random projections. Introductions to above mentioned feature space transformation techniques have been given in Section 2.3.

Eigenfaces (PCA) The famous Eigenfaces approach, introduced by Turk and Pentland in the early nineties [61, 311], is one of the best known approaches for face recognition. This method first reshapes each facial image into a column vector of gray-scale intensity values and then applies PCA to map the facial image vectors into a lower dimensional space. PCA aims to extract a subspace where the variance is maximized while the global structure of the image space is preserved. The output set of principle vectors is an orthonormal set of vectors representing the eigenvectors of the sample covariance matrix associated with the $m \ll n$ largest eigenvalues.

Fisherfaces (LDA) While the Eigenfaces method tries to preserve the global structure of the image space, the objective of LDA is to find the directions that are efficient for discrimination of classes. The Fisherfaces method was first introduced by Belhumeur *et al.* in [62]. Again, each gray-scale facial image is first vectorized into a column vector and then LDA is applied for subspace projection. LDA solves the Fisher criterion, i.e. the projection is chosen which maximizes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples. In other words LDA tries to find a subspace where the intra-class variance is minimized while the inter-class variance is maximized. The Fisherfaces method avoids the problem of singularity of the within-class scatter matrix by projecting the image set to a lower dimensional space using PCA before applying LDA.

Laplacianfaces (LPP) The LPP approach assumes that the face images reside on a nonlinear submanifold hidden in the image space. Unlike the Eigenfaces or Fisherfaces method, which effectively only see the global euclidean structure, LPP finds an embedding that preserves local information and obtains a subspace that best detects the essential face manifold structure. To preserve the local structure of the face space, this manifold structure is modeled by a nearest-neighbor graph. LPP will try to optimally preserve this graph when choosing projections. After constructing the graph, weights have to be assigned to the edges. Therefore a sparse symmetric

matrix which holds the weights of the edge joining vertices is constructed. LPP tries to find an embedding where samples that are close in the original feature space will stay close together in the smaller dimensional subspace. On the other hand, samples that were far apart in the original space will be mapped far apart in the new feature space. Similar to the Fisherfaces approach, the image set is usually projected into the PCA subspace before applying LPP by deleting the smallest principle components. Details about the algorithm and the underlying theory as well as the mathematical justification can be found in [58, 69].

The original implementations of Eigenfaces, Fisherfaces, and Laplacianfaces use a simple nearest neighbor classification method to obtain the final result. However, in general any classification algorithm can be used for identification. In the recent past a new and powerful classification methodology called Sparse Representation Classification (SRC) which is based on Compressed Sensing (CS), a technique for signal measurement and reconstruction, has been proposed.

Randomfaces Recently, SRC has been successfully applied to face recognition and promising results were obtained even under difficult lighting conditions and partial occlusion [71]. It is assumed, that all training samples of a single class lie on one mutual subspace. Thus, given a sufficiently large number of training samples, a testvector of a certain class can be represented as a linear combination of training samples of the same class. An introduction to SRC was given in Section 2.4.2. Usually, the high dimensional face images are first projected into a lower dimensional subspace using a sensing matrix which underlies the so called Restricted Isometry Property (RIP) [312]. It can be shown that even a randomly generated projection matrix can be used for that purpose. Such a matrix can simply be generated by sampling zero-mean independent identically distributed gaussian entries. A detailed description of the CS based face recognition algorithm and SRC in general can be found in [71].

After the huge success of SRC several adaptions and extensions have been proposed in the literature. Two of them are Robust Sparse Coding (RSC) and Regularized Robust Coding (RRC) which are reviewed in the subsequent paragraphs.

Robust Sparse Coding (RSC) The sparse coding model of SRC assumes that the coding residual follows a Gaussian or Laplacian distribution. In practice, however, such an assumption is often not accurate enough to describe the coding errors. In [313], Yang *et al.* proposed an algorithm called RSC that overcomes these limitations by modeling the sparse coding as a sparsity-constrained robust regression problem. RSC seeks for the Maximum Likelihood Estimation (MLE) solution of the sparse coding coefficients.

While in SRC the Probability Density Function (PDF) f_θ of the coding errors is explicitly taken as Laplacian or Gaussian distribution, one key-problem in RSC is the estimation of f_θ,

where θ is the parameter set that characterizes the distribution. Since it is hard to determine the PDF directly, the resulting modified version of the minimization problem is solved by an iteratively re-weighted sparse coding algorithm (IRSC). For a detailed explanation of the algorithm and the underlying theory the interested reader is referred to [313].

Regularized Robust Coding (RRC) The same authors later presented a generalized version of RSC named RRC, which could robustly regress a given signal with regularized regression coefficients [314]. By assuming that the coding residual and the coding coefficient are respectively independent and identically distributed, RRC seeks for a maximum a posteriori (MAP) solution of the coding problem. An iteratively reweighted regularized robust coding (IR^3C) algorithm is proposed to solve the RRC model efficiently. One important advantage of RRC is its robustness to various types of outliers (e.g., occlusion, corruption, expression, etc.) by seeking for an approximate MAP solution of the coding problem. By adaptively and iteratively assigning weights to the pixels according to their coding residuals, the IR^3C algorithm can robustly identify the outliers and reduce their effects on the coding process. The proposed RRC method was extensively evaluated for face recognition under different conditions, including variations of illumination, expression, occlusion, and image corruption. The experimental results clearly demonstrated that RRC outperforms previous state-of-the-art methods, such as SRC and RSC.

All the above mentioned methods use simple vectorized gray-level intensities as features. However, since simple pixel information is usually not robust enough to deal with challenging lighting conditions, partial occlusion, and other extrinsic factors, a number of more powerful and discriminating features has been proposed in recent years.

Gabor Sparse Representation Classification (GSRC) Gabor features are known to perform well in face and pattern recognition tasks for humans. Recently, Gabor features have been used in combination with the SRC scheme in [48] and outperformed the original Randomfaces algorithm [71], which uses basic pixel information as features. Yang *et al.* furthermore propose to apply PCA after feature extraction to obtain a more compact and discriminative representation of the feature space. The proposed GSRC approach was thoroughly evaluated on benchmark datasets for face recognition which include variations of illumination, expression, and pose, as well as block occlusion and disguise. The experimental results clearly demonstrated that the proposed GSRC has much better performance than SRC, leading to higher recognition rates at less computational costs.

Block-based Discrete Cosine Transform (BDCT) Ekenel *et al.* propose to represent the appearance of human facial regions using the 2D Discrete Cosine Transform (DCT) on non-overlapping blocks of the aligned face [315, 316, 307]. The region-based representation of the facial image provides robustness against appearance variation caused by facial expression and partial occlusion. The frequency information of the DCT on the other hand provides invariance to difficult lighting conditions. This is achieved by extracting the first ten AC coefficients of every block by zig-zag scanning. For lighting normalization, the first DCT coefficient which represents the DC (zero-frequency) component is discarded and extracted coefficients are subsequently normalized to unit norm. The overall feature vector is formed by concatenating the obtained coefficients for every block. Identification is done by a simple nearest neighbor distance-based classifier achieving state-of-the-art results on official benchmark datasets.

Local Binary Patterns (LBP) The main idea of LBP is to model the local texture of an image by using the joint distributions of differences between a center pixel and its surrounding neighbors. Hence, the original LBP descriptor, as proposed by Ojala *et al.* in [38], thresholds each 3×3 block within a monochrome image based on its center pixel to obtain a binary number for each block. As the neighborhood of each pixel consists of 8 neighbors, $2^8 = 256$ different labels can be obtained depending on the gray values of the center pixel and its eight neighbors. The statistics of these labels in form of histograms is then commonly used to model the local texture of images. The original LBP descriptor was later presented in a more general form by the same authors in [39] using different radii R and number of sampling points P. An introduction to LBP can be found in Section 2.2.1. Ahonen *et al.* were the first who applied LBP to the field of face recognition [317, 50]. As done in the BDCT approach [315] the face image is first divided into equally sized non-overlapping blocks. A LBP histogram is extracted for each region and the final feature vector is formed by concatenating the resulting histogram sequences. Ahonen *et al.* propose to use a simple χ^2-distance based nearest neighbor approach for classification.

Local Ternary Patterns (LTP) One known disadvantage of LBP, however, is that it may not work well for noisy images or flat regions such as cheeks or the forehead of human faces due to its thresholding paradigm which is solely based on the gray level value of the center pixel. Moreover, the reliability of LBP is known to decrease significantly for large illumination changes and shadowing. To overcome these limitations, Tan and Triggs proposed to replace LBP with a three-level operator called LTP in [40]. In LTP, the difference of the center pixel and its surrounding neighbors is encoded using three values (-1,0,1) according to a user specified threshold t. LTPs could then be encoded in a histogram of 3^P different bins, where P is the number of sampling points. However, this would result in extremely high dimensional feature vectors. Thus, in

practice LTPs are usually split into two different binary patterns, a "positive" and a "negative" part. Hence, for every LTP two LBPs can be obtained. The resulting two histogram sequences are concatenated in a subsequent step for the final LTP histogram descriptor. A more detailed review of LTP was given in Section 2.2.1. Other than that, Tan and Triggs use the same feature extraction pipeline as the LBP approach proposed by Ahonen *et al.* in [317, 50]. However, a sophisticated pre-processing pipeline including gamma-correction, Difference-of-Gaussian (DoG) filtering, and contrast normalization was proposed in [40] to enhance the systems robustness to difficult lighting conditions. Again, the χ^2-distance is used for a nearest neighbor classification paradigm to get a decision.

Local Gabor Binary Pattern Histogram Sequence (LGBPHS) In 2005, Zhang *et al.* successfully combined Gabor wavelets and LBPs to form a new face descriptor called LGBPHS [318]. First, so called Gabor Magnitude Pictures(GMPs) are obtained by convolving Gabor wavelets of different rotations and scales with the gray-scale input image. Each GMP is then divided into equally sized non-overlapping regions from where LBP histograms are extracted and concatenated to form the final face representation. For recognition, histogram intersection is used to measure the similarity of concatenated histograms and the nearest neighbor classification paradigm is applied to obtain a final decision. Experimental evaluations on publicly available benchmark datasets showed the effectiveness and robustness of the proposed approach

Although some of the above mentioned methods try to incorporate local information by applying features that efficiently describe the region within a local neighborhood or by dividing the facial image into non-overlapping blocks, the outcome of each of theses approaches is one single feature vector that is a representation of the global appearance of the face. However, recently also other techniques were proposed and evaluated that explicitly exploit local information. The most important and promising ones are briefly reviewed in the subsequent section.

Local Keypoint Methods

In contrast to holistic methods, face recognition techniques based on local keypoints first try to detect distinctive facial landmarks such as eyes, nose, and mouth to measure the geometric relationship between those fiducial points. Simplistic distance measures are subsequently exploited to match extracted geometrical features. Early approaches in the field of automatic face recognition were often based on these techniques. One of the earliest algorithms in this field date back to the mid-1970s [309], where relationships such as distances and angles between 16 different markings were used for recognition. More recently, Cox *et al.* reported a recognition rate of 95% on a dataset of 685 individuals in [319]. They manually annotated 35 different facial landmarks and computed a 30-dimensional feature vector based on a mixture of distances of

these points. However, facial landmarks were still annotated manually. Therefore, a significant decrease in accuracy can be expected for a completely automatic system where the location of automatically detected fiducial points is consequently not as accurate.

However, with the recent success of local keypoint detectors, robust and efficient algorithms based on local features were proposed. One of the most well-known methods is the Elastic Bunch Graph Matching (EBGM) approach by Wiskott *et al.* in [45] which is based on Dynamic Link Structures [320]. In EBGM, faces are represented as labeled graphs where the nodes represent local textures obtained by Gabor features (so called "jets") and the edges represent the distances between nodes. Hence, a face is represented as a collection of fiducial points and their spatial arrangement. All instances of frontal faces in the database are represented with the same kind of graph. A bunch graph is then created by combining the graphs of all faces in the database. Hence, a certain node of a bunch graph represents the texture of all variants of a specific facial landmark and the edges represent the mean distance between two fiducial points. More generally, a bunch graph is an abstract representation of object classes rather than of instances of a certain object. Thus, EBGM takes advantage of combinatorics of facial landmarks to represent a new face that was not seen before by the system. Recognition can then simply be done by comparing the graph of the new face to all graphs in the database and take the one with the highest similarity score. Albeit the fact that EBGM is one of the best performing algorithms for face recognition, extensive ground-truth annotation is necessary to properly train the algorithm. According to [321], elastic bunch graph matching only obtains reasonable results after manually placing the graphs for at least 70 facial training images per individual. Moreover, due to the complex 3D structure of a human face, the automatic placement of fiducial points becomes harder for off-frontal face images. However, a considerable amount of literature has been published proposing techniques to recognize faces from their profiles [322, 323, 324, 325, 326]. Another major drawback of EBGM is that it might not work well for data gathered from surveillance cameras due to the low-resolution character of the images and video sequences [300].

Hybrid Techniques

It is well known from psychophysics and neuroscience that both, holistic and local information, are crucial for perception and recognition of faces [327, 328, 329]. Thus, also a machine vision system should utilize both. One of the first hybrid approaches called *modular eigenfaces* was presented by Pentland *et al.* in [330] where the authors extended their earlier system [61] towards eigenfeatures such as eigeneyes, eigenmouth, etc. Experiments in [330] indicate that eigenfeatures extracted on different regions of the face are much more robust against different facial expressions than the holistic eigenfaces approach presented in [61]. This supports the assumption that locally extracted features are well suited for images with large variations.

Another interesting biologically inspired approach for hybrid face recognition called Local Feature Analysis (LFA) in combination with PCA was presented by Penev and Atick in [331]. Unlike the global eigenmodes, LFA gives a description of the face in terms of statistically derived local features and their positions. In [331], the authors successfully combine the global face representation extracted by PCA and the local information by LFA to enhance the recognition performance of both modalities alone.

Also flexible models such as Active Shape Models (ASMs) and Active Appearance Models (AAMs) which use both, shape and gray-level information, have been used to recognize faces [332, 333]. ASMs and AAMs are statistical models of generic objects that are deformable so they can fit themselves to the shape of an object in a new image. After the flexible appearance model was fitted to a new face, shape parameters as well as local gray-value information at each model point are collected. Then, the face image is transformed to a mean face shape and shape-free model parameters can be obtained. All three parameter-sets, i.e. shape parameters, local gray-value information at the model points, and the shape-free model parameters are used for classification.

A pose and illumination invariant face recognition approach which combines 3D morphable models and component-based face recognition techniques was presented by Huang *et al.* in [334]. First, a 3D morphable model of a face of every person in the database is constructed based on three face images in different poses (frontal, semi-profile, and profile). Once this model is constructed it can be used to generate arbitrary synthetic images of the same person in various poses and different lighting conditions. Then, a component-based face recognition system can be used to identify unseen test images. Similar to EBGM, the main idea of component-based methods is to decompose a face into its main components, e.g. eyes, mouth, and nose, and model the interconnections between them with a flexible geometrical model. However, in Huang *et al.* simple gray-scale components were used instead of Gabor features as in [45]. Although the proposed system achieved impressive results in experiments conducted in [334], one major drawback of the system is that the generation of the 3D model is person-specific and therefore requires cooperation in order to get high-quality images. However, recent success in face recognition based on 3D morphable models might lead to powerful and robust face recognition algorithms applicable in real-world environments [335, 336]

A lot of effort has also been put into face recognition using neural networks. One of the first applications of neural networks to the field of face recognition was presented by Lin *et al.* in [337] where they proposed a probabilistic decision-based neural network (PDBNN) for identification. The system was evaluated on two public benchmark datasets and achieved state-of-the-art results at that time. Later, a Radial Basis Function (RBF) neural classifier was used by [338] to cope with the problem of small training sets. Evaluation on publicly available datasets demon-

strated the efficiency of the proposed learning algorithm with regard to classification and learning efficiency. With the recent development and success of deep learning [339, 340, 341], artificial neural networks regained attraction for face recognition. Most recently, Taigman *et al.* [342] developed an algorithm called *DeepFace* which successfully combines 3D modeling of a human face for alignment and a nine-layer deep neural network for recognition. The system was trained on an extremely large dataset from *Facebook* consisting of four million facial images belonging to more than $4,000$ individuals. It was then tested on the Labeled Faces in the Wild (LFW) dataset which is one of the most widely acknowledged benchmark dataset for face verification and recognition in unconstrained environments, achieving an accuracy of 97.35% which is close to human-level performance.

Table 3.4 gives a brief summary of the face recognition approaches discussed above.

Category	Reference	Method	Description	Notes
Holistic Methods	Turk *et al.* [311]	Eigenfaces	(1) Column vectors of gray-scale information as features (2) Dimensionality reduction using PCA (3) Euclidean distance-based nearest neighbor classifier	(1) Fast and easy to implement (2) Gray-scale information not robust enough for recognition in real-world environments
	Belhumeur *et al.* [62]	Fisherfaces	(1) Column vectors of gray-scale information as features (2) Dimensionality reduction using PCA followed by LDA (3) Euclidean distance-based nearest neighbor classifier	(1) Fast and easy to implement (2) Outperforms Eigenfaces approach on most datasets (3) Takes class affiliation of training data into account (4) Limited performance in real-world settings due to simple gray-scale information as features

Category	Reference	Method	Description	Notes
	He *et al.* [58]	Laplacian-faces	(1) Column vectors of gray-scale information as features (2) Dimensionality reduction using PCA followed by LPP (3) Euclidean distance-based nearest neighbor classifier	(1) Fast and easy to implement (2) Assumes that face images reside on a non-linear submanifold (3) Outperforms Eigenfaces and Fisherfaces approach (4) Simple gray-scale information not robust enough in real-world settings
	Wright *et al.* [71]	Random-faces	(1) Column vectors of gray-scale information as features (2) Dimensionality reduction using random projection (3) CS-based classification (SRC)	(1) Good results even under difficult lighting condition & partial occlusion (2) Assumes that test image can be represented as linear combination of training images of particular class
	Yang *et al.* [313]	Robust Sparse Coding (RSC)	(1) Column vectors of gray-scale information as features (2) SRC assumes that coding errors follow Gaussian or Laplacian distribution (3) RSC overcomes this limitation of standard SRC by iteratively estimating PDF of coding errors	(1) Outperforms Randomfaces on official benchmark datasets (2) Slower and computationally more expensive than Randomfaces

Category	Reference	Method	Description	Notes
	Yang *et al.* [314]	Regularized Robust Coding (RRC)	(1) Column vectors of gray-scale intensities as features (2) Generalization of RSC (3) RRC robustly identifies outlier pixels and reduce their effect on coding errors	(1) Robust to various outliers caused by expression, partial occlusion, and illumination changes (2) Outperforms Randomfaces and RSC (3) Slower and computationally more expensive than Randomfaces and RSC
	Yang *et al.* [293]	Gabor Sparse Representation Classification (GSRC)	(1) Gabor-based features as descriptors (2) PCA for dimensionality reduction (3) SRC for classification	(1) Robust to partial occlusion and illumination variation (2) Outperforms Randomfaces approach (3) Classification computationally less expensive than Randomfaces due to smaller feature size
	Ekenel *et al.* [308]	Block-based Discrete Cosine Transform (BDCT)	(1) Concatenation of first 10 AC DCT-coefficients on non-overlapping blocks (2) Classification based on nearest neighbor approach using normalized correlation-based distance measure	(1) State-of-the-art results on benchmark datasets (2) Robustness to different lighting conditions by omitting DC component of 2D-DCT (3) Robust to partial occlusion due to block-based processing
	Ahonen *et al.* [50]	Local Binary Patterns (LBP)	(1) Face image divided into equally sized non-overlapping blocks (2) Extraction of LBP histograms on 3×3 neighborhood (3) χ^2-distance based nearest neighbor classifier	(1) Outperforms pixel-based face recognition approaches (2) Fast and easy to implement (3) LBP known to be prone to noise in non-textured regions and illumination changes

Category	Reference	Method	Description	Notes
	Tan *et al.* [40]	Local Ternary Patterns (LTP)	(1) Extension of LBP approach (2) Difference of center pixel & surrounding neighbors coded by three-level operator [-1,0,1] (3) LTP split into "positive" and "negative" part before building histograms (4) Nearest neighbor classification using χ^2-distance	(1) Performs better than LBP-based face recognition (2) Additional preprocessing stages proposed to increase robustness of descriptors (3) Higher-dimensional feature vectors than LBP, thus feature space transformation might be necessary for small sample sizes
	Zhang *et al.* [318]	Local Gabor Binary Pattern Histogram Sequence (LGBPHS)	(1) Descriptor based on combination of Gabor features & LBP (2) Block-based LBP-histograms extracted on every Gabor Magnitude Picture (GMP) (3) Nearest neighbor classification based on histogram intersections	(1) Evaluation on benchmark datasets shows superiority of basic LBP approach (2) Combination of Gabor features and LBP is more robust than both features alone
Local Keypoint Methods	Cox *et al.* [319]	Manual landmarks	(1) Manual annotation of 35 landmark points (2) Extraction of 30-D feature vector based on mixture of distances	(1) High accuracy on benchmark datasets (2) Facial landmark points have to be localized manually (3) Decrease in performance can be expected for a full-automatic recognition system

Category	Reference	Method	Description	Notes
	Wiskott *et al.* [45]	Elastic Bunch Graph Matching (EBGM)	(1) Faces are represented by labeled graphs (collection of fiducial points & their arrangement) (2) Nodes of graphs obtained by Gabor features ("jets"), edges represented by distances between jets (3) Bunch graph: combination of graphs of all faces of an individual (4) Classification by distance of test graph with all graphs in the database	(1) One of the best performing face recognition algorithms (2) Extensive ground-truth annotation necessary (at least 70 training images per class [321]) (3) Not well suited for low-resolution data due to automatic detection of interest points
Hybrid Techniques	Pentland *et al.* [330]	Modular Eigenfaces	(1) Extension of Eigenfaces approach to Eigenfeatures (e.g. eigeneyes, eigennose, eigenmouth, etc.) (2) PCA on local regions instead of whole face (3) classification of local regions and majority voting	(1) Experiments indicate that Eigenfeatures extracted on local regions are more robust against local changes than global Eigenfaces approach (2) Face regions represented by gray-level values (3) Gray-scale information not robust enough for face recognition in natural environments
	Penev *et al.* [331]	Local Feature Analysis (LFA)	(1) Description of faces in terms of statistically derived local features & their positions (2) Combination of global Eigenfaces approach and local information obtained by LFA	(1) Combination of global face representation & local features enhances accuracy and robustness of the system (2) Gray-scale information of global approach might be not robust enough

Category	Reference	Method	Description	Notes
	Lanitis & Cootes *et al.* [332, 333]	Active Shape Models (ASMs) Active Appearance Models (AAMs)	(1) Statistical models of generic deformable objects (2) Flexible appearance models are fit to face and are subsequently normalized (3) Shape-free model parameters, gray-level information at keypoints, and shape parameters used for classification	(1) Good performances on various benchmark datasets (2) Computationally expensive (3) Extensive ground-truth annotation necessary (4) Hard to automatically localize facial landmark points in real-world environments
	Huang *et al.* [334]	3D morphable models	(1) Generation of 3D morphable model of every person (frontal, semi-profile, profile) (2) 3D model used to generate arbitrary synthetic face images (3) Recognition similar to EBGM but with gray-scale information as features	(1) Impressive results on benchmark datasets (2) 3D models are person-specific, i.e. algorithm requires close cooperation with individuals in order to collect high-quality data
	Taigman *et al.* [342]	DeepFace	(1) Generic 3D model for face alignment (2) Deep neural network for recognition	(1) Excellent results on Labeled Faces in the Wild (LFW) database (2) Results comparable with human performance (3) Huge amounts of training data necessary to train deep neural network

Table 3.4.: ***Overview of state-of-the-art algorithms for automatic face recognition.***

Although face recognition algorithms such as EBGM, AAM or methods based on 3D morphable models are capable to reliably identify human beings, they usually require high-resolution images or videos and often even cooperation of the subject to be identified. Hence, these methods are not suited well for an application as presented in this thesis since one requirement of the system is to reliably recognize individuals even under challenging lighting conditions without any external intervention. Impressive results were recently achieved by the *DeepFace* algorithm proposed by Taigman *et al.* in [342]. However, collecting and annotating huge amounts of data in natural habitats of great apes is often infeasible which makes training of deep neural network-based approaches impractical for the task at hand. Keeping all these factors in mind, a combination of holistic face recognition approaches and local keypoint methods is used within the proposed PRF which is assumed to be the best choice considering the challenging conditions in which images and videos of primates are usually captured.

3.4. Chapter Summary

In this chapter, an overview of existing computer vision approaches for non-invasive animal monitoring was given. State-of-the-art algorithms in the new and growing research discipline of *Visual Animal Biometrics* were reviewed and advantages as well as disadvantages of the proposed approaches were revealed. It was shown that a number of promising algorithms are already available to support biologist with tedious annotation work of remotely gathered image and video footage. This not only includes the detection, tracking, and behavioral analysis of animals (see Section 3.2.1) but also individual identification and species recognition (see Section 3.2.2). However, a vast number of proposed algorithms was evaluated on datasets gathered in constraint environments that do not reflect difficulties present in real-world scenarios. Although a minority of systems exist that have proven to achieve adequate results even in natural settings, they are often limited to patterned animals and utilize individually unique markings on fur, skin or other external organs for detection or identification. However, for a variety of non-patterned animal species such approaches are often infeasible since unique markings are either not present or cannot be used in a non-intrusive way due to limited resolution of gathered image and video material. Furthermore, since the primate identification system presented in this thesis is based on face recognition, the second part of this chapter focused on state-of-the-art face recognition algorithms developed to identify humans in images or video sequences. After presenting a coarse taxonomy of face recognition techniques, a number of holistic algorithms were reviewed more carefully since they serve as benchmark methods for the proposed PRF. A thorough evaluation and comparison of these algorithms applied to the datasets used within this thesis is given in Chapter 5.

4. Identification of Primates using Face Recognition

4.1. Chapter Overview

As already discussed in Chapter 3.2, most state-of-the-art algorithms for automated non-invasive identification of animals in images or videos are based on characteristic coat patterns or other individually unique natural markings. Unfortunately, such approaches are often infeasible for the identification of great apes since unique patterns on fur or skin are not existent or cannot be used due to limited camera resolution. Based on the assumption that humans and our closest relatives share similar properties of their facial appearance, a completely autonomous unified face recognition framework, including face detection, face alignment, and face recognition, is proposed to identify primates in their natural habitats. A detailed description of the proposed facial recognition system for images and videos is given in this chapter.

Section 4.2 introduces the proposed Primate Recognition Framework (PRF) for recognition of primates in still images. It is first explained how faces as well as facial features such as eyes and mouth are automatically detected. Secondly, based on the located eye and mouth coordinates detected faces are aligned in order to achieve comparability of faces across the entire database. Finally, global as well as local visual descriptors are extracted from the aligned and normalized faces in order to subsequently recognize them. To further enhance the system's robustness to various extrinsic and intrinsic factors a simple but efficient decision fusion scheme is proposed which combines the advantages of global face representation as well as information extracted from local keypoints.

Section 4.3 extends these ideas and shows how temporal information in video recordings can be efficiently exploited to further enhance the system's performance. Therefore, once a face is detected it is tracked through the video sequence. Furthermore, a number of quality assessment modules are proposed in order automatically select the frames which are best suited for recognition. Hence, identification is done only in those frames of a given face-track which contain high-quality facial images. Finally, a novel frame-weighting technique assigns weights to each processed frame and aggregates the results in order to obtain a final prediction per face-track.

4.2. Face Recognition in Images

Figure 4.1 gives an overview of the proposed system for primate face recognition in images. It comprises three main components:

1. **Face and facial feature detection**: First, primate faces are automatically detected. This step not only includes the location of faces but also an estimate of size and resolution. Furthermore, facial features, i.e. both eyes and the mouth, are located automatically within each face region.

2. **Pre-processing**: Secondly, a number of pre-processing steps such as face alignment and gray-scale conversion are applied to ensure comparability of facial images.

3. **Face Recognition**: The third and last step recognizes the detected and normalized faces and assigns identities to them if they are known to the system. However, if the detected face belongs to an individual which is not represented in the training database it gets rejected as unknown.

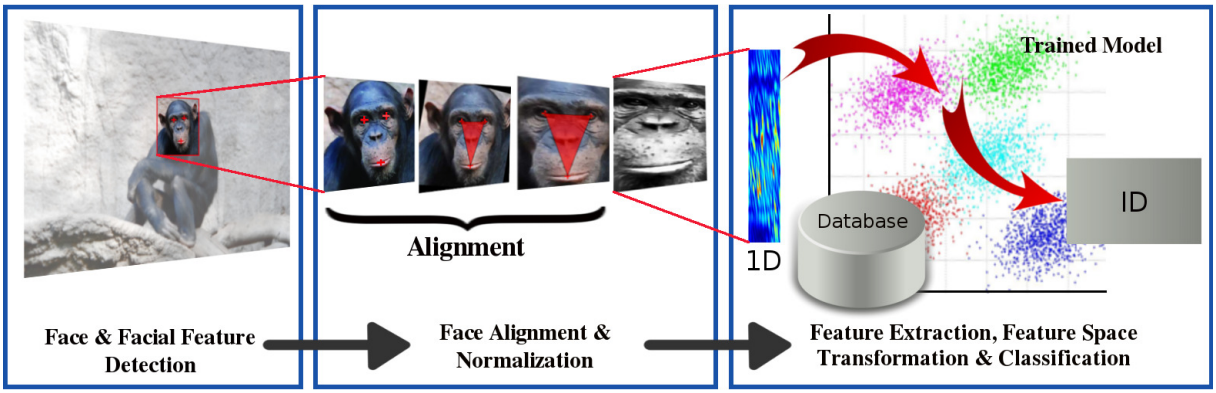


Figure 4.1.: ***Overview of the proposed system for primate face identification in images.*** After all possible faces in an image or video sequence were detected, each face is aligned and converted to gray-scale in order to ensure comparability of faces across the entire dataset. In the final stage of the proposed framework, the detected and aligned faces are identified using feature extraction, feature space transformation and classification techniques.

A detailed description of every component of the proposed system is given in the following subsections. Although parts of the developed PRF were already published in various scientific conferences and journals [2, 3, 4, 5, 6, 7, 8, 9, 10, 11] each module necessary for accurate and robust detection and identification of great apes based on their facial appearance is covered in more detail within this thesis.

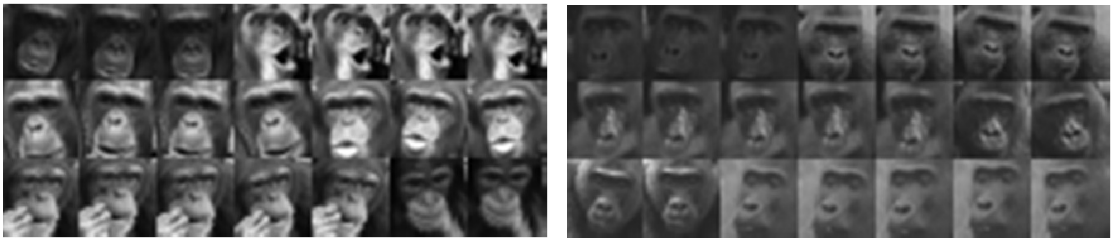
4.2.1. Face and Facial Feature Detection

Within this thesis an external face detection library called Sophisticated High-Speed Object Recognition Engine (SHORE™)[1], developed by Fraunhofer Institute for Integrated Circuits (IIS), is utilized to localize faces of great apes in image and video footage. SHORE™ allows real-time robust detection and tracking of frontal primate faces in images and videos. For completeness, a brief overview of the face detection library is given in this section. For a more thorough and detailed description of the algorithms used for face detection and tracking, the reader is referred to [9, 113, 343].

Rigid object detectors are known to perform well for human face detection in images and videos [99]. Starting from the assumption that humans and our closest relatives share similar properties of the face, state-of-the-art face detection algorithms should therefore be able to detect faces of great apes as well. For this purpose, SHORE™ has been extended to automatically localize faces of primates. Ernst and Küblbeck [113] utilize a detection model comprising multiple consecutive classification stages with increasing complexity to detect faces of chimpanzees and gorillas in images and video sequences. Each stage comprises a feature extraction step and a look-up table based classifier built in an offline training procedure using Real-AdaBoost [115].

Real-time capability is achieved by using simple and fast pixel-based features in the first stages and more sophisticated and therefore more complex descriptors in subsequent stages. The first stages can be considered as a fast but inaccurate candidate search while the remaining stages focus on slower but more accurate classification. Each stage comprises one out of three illumination invariant features: *edge orientation features*, *census features*, and *structure features*. Edge orientation features represent pixel-based gradient directions and are extracted by first applying a 3×3 Sobel operator in horizontal and vertical direction which results in gradient images I_x and I_y. The final feature vector is constructed by calculating quantized gradient directions using $atan2(I_x, I_y)$ and a quantization interval of 35 bins. In subsequent classification stages more complex but also more sophisticated features called census features [114] are extracted which encode local brightness changes within a 3×3 neighborhood around each pixel. Therefore, each pixel is compared to its eight neighbors and an 8-digit binary number is derived by assigning a "1" if the intensity of the neighborhood pixel is higher than the current pixel and "0" otherwise. In the final classification stages, structure features which are built out of scaled versions of census features are extracted on image regions. To improve the system's robustness against facial variations and in-plane rotations, a set of annotated face samples are cut out and modified with slight random variations such as rotation, mirroring, and translation. Figure 4.2 shows positive examples of the face detection training set for chimpanzees (a) and gorillas (b).

[1] `http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore.html` Last visit: January 23^{rd}, 2014



(a) (b)

Figure 4.2.: ***Positive examples of chimpanzee and gorilla faces used to train the face detector.*** Annotated chimpanzee (a) and gorilla faces (b) were cut out multiple times and modified with slight random modifications in order to increase the systems robustness to facial variations. Image source: [113].

Non-face data was generated by randomly cropping patches from images without faces. This data serves as initial negative training data. Subsequently, further non-face data was gathered by bootstrapping the initial model on images without ape faces.

For the actual detection, the gray scaled input image is initially convolved with a 3×3 mean filter kernel to compensate noise. While the detection model is fixed with a size of 24×24 pixels, the mean filtered image is downscaled multiple times using a scaling factor of 1.24 to build an image pyramid. A real-time capable coarse to fine search is applied by shifting the detection window across every pyramid level to achieve scale invariance. To ensure real-time performance, only candidate face regions which achieve high confidences in the first stages reach the slower but more accurate detection stages. Detections in multiple pyramid levels are subsequently merged to a single detection with mean size and location. Examples of detected primate faces in their natural habitats can be seen in Figure 4.3. Although cluttered background and extreme lighting conditions place high demands on the face detection framework, SHORE™ provides accurate results for frontal face detection with low false-positive and negative rates. A thorough evaluation of the detection performance can be found in [9, 113].

After face detection, SHORE™ additionally locates facial features such as eyes and mouth within each face region using the same algorithms as explained above. However, detection models for facial features are simpler and less powerful compared to the original facial models since few false positive detections can be expected within a small region of interest. Eye and mouth regions were cut out from annotated training samples and detection models of size 16×16 pixels were generated for training. Within each located face region, predefined areas around both eyes and the mouth are scanned thoroughly by applying the appropriate model in different pyramid levels to locate facial feature points. In case eyes or mouth could not be detected, fixed markers relative to the borders of the region of interest are used as coarse facial feature locations.

Figure 4.3.: ***Detection results by SHORE™ of near-frontal chimpanzee and gorilla faces in their natural habitats.*** Detected faces are marked with green rectangles. The species is automatically assigned to every face using the detection scores of the chimpanzee and the gorilla model, respectively. Both scores are superimposed in yellow ([*chimp-score*|*gorilla-score*]). In case one detection model did not find a face, the corresponding detection score is replaced by hyphens. Although background clutter and challenging lighting conditions place high demands on SHORE™ only few faces were missed (blue arrow) and few false positive detections occurred (red arrows). Image source: [113]. [I02, I03]

In addition to real-time capable face detection in images and videos, Ernst *et al.* also proposed methods to automatically distinguish between chimpanzees and gorillas [113]. The first approach utilizes the detection scores of the applied face detection models for chimpanzees and gorillas. For the second method a separate classification model is trained based on structure features only and applied to the detected face. Both techniques for species classification perform remarkably well with over 90% accuracy. Although SHORE™ has been proven to be robust against difficult lighting situations, it lacks in robustness to severe occlusion and far-off frontal poses. As demonstrated by Sandwell and Burghardt in [126], Deformable Part Based Models (DPM) can be used to locate primate faces in non-frontal poses as well. However, as stated in [126], this approach is not yet real-time capable. Furthermore, faces often are expected to be in a near-frontal pose for subsequent analysis such as individual identification for instance. Therefore, the approach by Ernst and Küblbeck [113] has been used in this thesis to automatically detect faces and facial features of chimpanzees and gorillas.

4.2.2. Face Alignment

An important step for robust and accurate face recognition is the alignment of detected faces. An *affine transformation* is applied to ensure that facial features such as eyes and mouth are located at the same positions throughout the entire dataset which guarantees comparability of extracted visual descriptors for all faces.

Moreover, a proper alignment of faces minimizes the influence of unwanted background which significantly improves recognition performance.

A thorough introduction to projective geometry and transformations in 2D space can be found in [344]. As defined in [344], in general a projective transformation is a linear transformation of a homogeneous 3D-vector $\boldsymbol{z} = [x, y, z]^{\mathrm{T}}$ represented by a 3×3 matrix $\boldsymbol{H}$ as

$$\boldsymbol{z}' = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \boldsymbol{H}\boldsymbol{z}, \tag{4.1}$$

were $\boldsymbol{H}$ is a non-singular transformation matrix and $\boldsymbol{z}' = [x', y', z']^{\mathrm{T}}$ is the transformed vector. It can be shown that projective transformations form a group which can be further divided into several specializations or subgroups: the *Euclidean* group, the *Similarity* group, and the *Affine* group which form a hierarchy of transformations.

As stated earlier, an affine transformation is used in this thesis for face alignment. It can be shown that an affine transformation actually is a composition of two fundamental transformations: *rotation* and *anisotropic scaling* [344]. Thus, Equation 4.1 reduces to

$$\boldsymbol{z}' = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{t} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \boldsymbol{z} = \boldsymbol{H}_{\mathrm{A}}\boldsymbol{z} \tag{4.2}$$

where $\boldsymbol{A}$ is a 2×2 non-singular matrix and $\boldsymbol{t}$ is a 2D column vector. In order to compute the transformation matrix $\boldsymbol{H}_{\mathrm{A}}$, three point correspondences are required. Hence, the coordinates of the three facial feature points detected by SHORE™ are utilized for face alignment: the left eye, the right eye, and the mouth (see Section 4.2.1). Thus, the first step of face alignment is to find a transformation matrix $\boldsymbol{H}_{\mathrm{A}}$ which maps these landmark points to predefined coordinates. Once $\boldsymbol{H}_{\mathrm{A}}$ has been computed it can be applied to the whole face image for alignment. Within the proposed application, the automatically detected landmark points are mapped as follows:

$$\begin{aligned} \text{Left eye:} \quad & \boldsymbol{x}_{\mathrm{LE}} = [x_{\mathrm{LE}}, y_{\mathrm{LE}}]^{\mathrm{T}} \quad \mapsto \quad [0.75w, 0.3h]^{\mathrm{T}} \\ \text{Right eye:} \quad & \boldsymbol{x}_{\mathrm{RE}} = [x_{\mathrm{RE}}, y_{\mathrm{RE}}]^{\mathrm{T}} \quad \mapsto \quad [0.25w, 0.3h]^{\mathrm{T}} \\ \text{Mouth:} \quad & \boldsymbol{x}_{\mathrm{M}} = [x_{\mathrm{M}}, y_{\mathrm{M}}]^{\mathrm{T}} \quad \mapsto \quad [0.5w, 0.9h]^{\mathrm{T}}, \end{aligned}$$

where w and h are the width and the height of the cropped face region, respectively. Figure 4.4 illustrates the proposed face alignment procedure on a chimpanzee face.

It can be seen in Figure 4.4 that the applied face alignment step not only ensures that facial features are located at the same position throughout the entire dataset but also minimizes the effect of background clutter which is a prerequisite to perform robust face recognition in practice.

(a) (b) (c)

Figure 4.4.: ***Face alignment of a chimpanzee face.*** The figure illustrates the proposed face alignment procedure using an affine transform. The coordinates of the left and right eye as well as the mouth of the unaligned image (a) are used the compute a transformation matrix $\boldsymbol{H}_{\mathrm{A}}$. The face is then rotated into an upright position such that both eyes lie on a horizontal line (b). Finally, the face is warped such that both eyes and the mouth are located at the same predefined positions throughout the entire dataset (c). Note that these transformations are actually performed in one single step by applying the obtained affine transformation matrix $\boldsymbol{H}_{\mathrm{A}}$.

One of the main purposes of the proposed PRF is to support the annotation of camera trap data recorded by biologists. Hence, a basic requirement of the developed system is its independence of the recording type, i.e. face detection and recognition should work equally well for color, gray-scale, or infrared image and video footage. Therefore, the aligned face images are subsequently converted to gray-scale. Moreover, all facial images are then scaled to a fixed size of 64×64 pixels which is also the minimum image sized allowed to be processed within the proposed framework. The choice of the minimal image size is critical: Too low resolved face images lack in facial details which are important for accurate identification. Too large image sizes on the other hand would result in high false negative rates for low-resolution footage since many faces would not be processed by the system. Typical choices for critical image sizes in the face recognition literature are typically 128×128 pixels and above. Wright *et al.* for instance chose a minimal image size of 192×168 pixels for their experiments [71], whereas Ahonen *et al.* observed that their algorithm performs best with an image size of at least 128×128 pixels. Keeping in mind the relatively low-resolution video files obtained from camera traps commonly used by biologists during field studies, a size of 64×64 pixels is an acceptable trade-off between face recognition performance, processing time, and false negative rates.

It will be shown in Section 5.4.1 that the face descriptors used within the proposed framework do not benefit from an additional lighting normalization. Therefore, no further pre-processing and normalization steps are performed after face alignment, resizing, and gray-scale conversion.

4.2.3. Individual Identification

Individual identification is the main part of the proposed system and comprises three steps: *feature extraction*, *feature space transformation*, and *classification*. First, global as well as local visual features are extracted which are both well suited for discrimination. While global descriptors represent the whole appearance of an ape's face, local features around certain facial fiducial points are more robust against local changes as they only encode detailed traits of the corresponding point of interest. Furthermore, it is well known from psychophysics and neuroscience that both holistic and local information are crucial for perception and recognition of faces [327, 328, 329].

As both descriptors are too high-dimensional to perform fast and efficient face recognition in practice, a feature space transformation technique called Locality Preserving Projections (LPP) [58] is subsequently applied to obtain a lower dimensional feature space without losing information which is important for discrimination. The resulting lower dimensional feature vectors are then used for classification. After classifying global and local feature vectors separately, a decision fusion technique that takes the confidences of both classifiers into account is applied which yields the final result. Figure 4.5 illustrates the proposed decision fusion scheme.

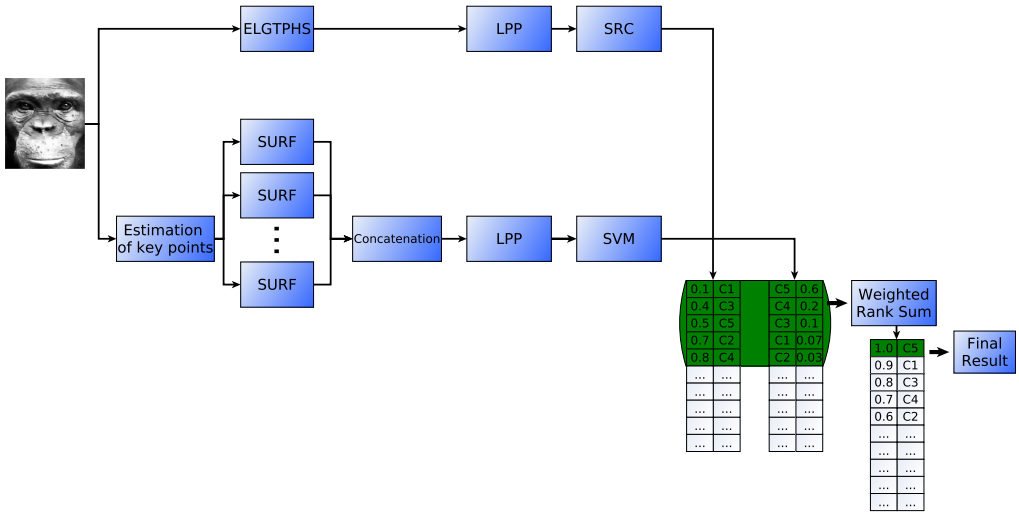


Figure 4.5.: ***The proposed decision fusion scheme to identify primates.*** The figure shows the proposed fusion scheme which combines the results of global Gabor-based features and local Speeded-Up Robust Features (SURF). Local features are extracted from six different keypoints around both eyes and the nose which are subsequently concatenated. Both global and local features are then projected into a smaller dimensional subspace using LPP. The global descriptor is classified using Sparse Representation Classification (SRC) while the local feature vector is classified by a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel. The ranked results are then combined using a decision fusion rank sum method in order to obtain the final result.

4.2.3.1. Face Recognition Using Global Features

Global Face Representation using Enhanced Local Gabor Ternary Pattern Histogram Sequence (ELGTPHS) Inspired by the work of Zhang *et al.* [318] a robust face descriptor called ELGTPHS is proposed in this section to compactly capture the global appearance of a primate's face. A descriptor of a facial image is extracted by concatenating the histograms resulting from Extended Local Ternary Patterns (ELTP) [40, 345] over different regions of the image. To further enhance the discriminative power of the descriptor, instead of extracting ELTPs directly on the gray scale intensity image they are computed on several Gabor Magnitude Pictures(GMPs) obtained from convolving the facial image with multi-scale and multi-orientation Gabor kernels [37]. Thus, two complimentary descriptors are combined on the feature level due to ELTP's capability to capture small appearance details while Gabor wavelets encode facial features on a broad range of scales and orientations. Furthermore, spatial information is encoded by dividing each GMP into multiple non-overlapping cells and calculating a histogram representation for each block separately before concatenation.

Figure 4.6 illustrates the process of extracting the proposed ELGTPHS descriptor on facial images, while details are given in the subsequent sections.

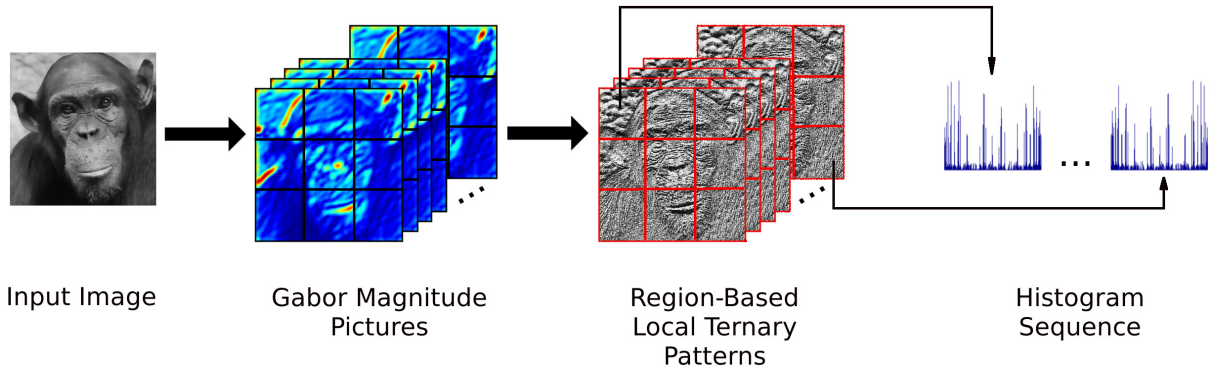


Figure 4.6.: ***The proposed global feature extraction pipeline.*** The figure shows the process of extracting the proposed ELGTPHS descriptor. First, GMPs are calculated by convolving the gray-scale input image with Gabor kernels of different orientation and scales. Subsequently, each GMP is divided into non-overlapping blocks and ELTP histograms are extracted from every region separately. The obtained histograms from each region of all GMPs are then concatenated to form the final feature vector.

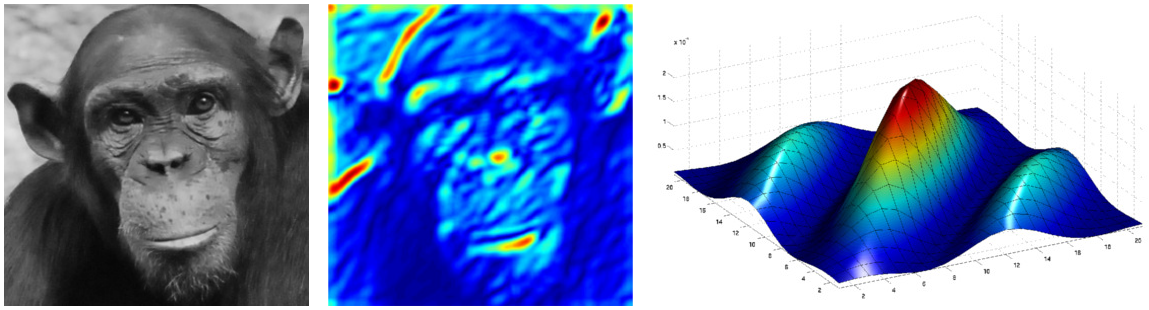
Conclusively, the calculation of the global facial descriptor used in this thesis can be summarized as follows:

1. GMPs are calculated by convolving Gabor wavelets at different orientations and scales with the (aligned) gray-scale input image.
2. Each GMP is divided into multiple non-overlapping regions to preserve spatial information.
3. ELTP histograms are extracted from each cell of all GMPs and concatenated to form the final feature vector.

Gabor Magnitude Pictures(GMPs) The first step to extract the proposed ELGTPHS descriptor is to calculate the GMPs by convolving the input image $\boldsymbol{I}(z)$ with a set of 2D Gabor kernels $\boldsymbol{\Psi}_{\mu,\nu}(z)$, where $z = (x, y)$ is the pixel location.

$$\boldsymbol{G}_{\mu,\nu}(z) = \boldsymbol{I}(z) * \boldsymbol{\Psi}_{\mu,\nu}(z) \tag{4.3}$$

An introduction to Gabor wavelets was given in Section 2.2.1. Figure 4.7 shows the GMP as a result of the convolution of an input image and a 2D Gabor kernel.



(a) (b) (c)

Figure 4.7.: ***The Gabor Magnitude Picture (GMP) and the according Gabor wavelet.*** The figure shows the result of the convolution of an input image (a) with a Gabor kernel. The resulting GMP is illustrated in (b) while the real part of the according Gabor kernel of size 21×21 is shown in (c). The parameters for the generation of the Gabor kernel were set as follows: $k_{max} = \frac{\pi}{2}$, $f = \sqrt{2}$, $\sigma = \pi$, $\theta_\mu = \frac{3\pi}{4}$, and $\nu = 3$.

As in most image processing applications a square kernel with odd side-length H is used in this thesis. In order to obtain a suitable kernel size, the wavelets are further cropped such that p percent of the energy of all Gabor wavelets are preserved, where p is a user-defined parameter. Let $|\boldsymbol{\Psi}_{\mu,\nu}(z)|$ be the magnitude of the complex Gabor wavelet at rotation μ and scale ν be defined as

$$|\boldsymbol{\Psi}_{\mu,\nu}(z)| = \sqrt{\Re\{\boldsymbol{\Psi}_{\mu,\nu}(z)\}^2 + \Im\{\boldsymbol{\Psi}_{\mu,\nu}(z)\}^2}, \tag{4.4}$$

where $\Re\{\cdot\}$ denotes the real part and $\Im\{\cdot\}$ the imaginary part of the complex Gabor kernel. The energy ratio over all $F \cdot S$ Gabor kernels $\mathcal{R}(s)$ can then be calculated as

$$\mathcal{R}(s) = \sum_{\nu=0}^{S-1} \sum_{\mu=0}^{F-1} \left(\frac{E_{\mu,\nu}}{E^s_{\mu,\nu}} \right), \tag{4.5}$$

where the total wavelet energy is defined as $E_{\mu,\nu} = \sum_z |\boldsymbol{\Psi}_{\mu,\nu}(z)|^2$ and $E^s_{\mu,\nu} = \sum_{z=s}^{H-s} |\boldsymbol{\Psi}_{\mu,\nu}(z)|^2$ is the energy of the cropped version of the Gabor kernel at scale ν and rotation μ. The final index s_f is then given by

$$s_f = s_c - 1 \quad \text{where the condition} \quad \mathcal{R}(s_c) < \frac{p}{100} \quad \text{is reached.} \tag{4.6}$$

As proposed in [48], p was set to 90 within the proposed framework which means that all Gabor wavelets were cropped such that 90% of their overall energy are preserved within the filter kernels.

After convolving the input image $\boldsymbol{I}(z)$ with each Gabor wavelet obtained from all scales and all orientations, the set $\boldsymbol{\mathcal{S}} = \{|\boldsymbol{G}_{\mu,\nu}(z)| : \nu \in \{0, \ldots, S\}, \mu \in \{0, \ldots, F\}\}$ forms the overall GMP representation of the image $\boldsymbol{I}(z)$ from where region-based ELTP histograms are extracted. A common choice for the number of orientations F and the number of scales S for the creation of Gabor wavelets is 8 and 5, respectively, which is agreed to perform well in most pattern recognition tasks and face recognition in particular [48, 42, 346]. Setting the number of orientations to 8 corresponds to an angle of $\frac{\pi}{4}$ between consecutive wavelets which means that edge information of 45° increments are represented in the final descriptor. A total number of 5 scales is a good choice for most face recognition applications since coarse structures as well as granular details of a face can be captured for identification. All other parameters for the generation of Gabor kernels are set as described in [48] and are summarized in Table B.1.

Extended Local Ternary Patterns (ELTP) As stated earlier, Gabor-based features and ELTP are fused on the feature level in order to combine the advantages of both descriptors. Region-based ELTP histograms are extracted from the multi-scale, multi-orientation GMPs obtained from the previous step and are subsequently concatenated to form the final feature vector.

In order to encode spatial information into the final descriptor, every GMP is first divided into $B \times B$ non-overlapping quadratic regions from where ELTP feature histograms are subsequently extracted. Too many number of blocks would result in unnecessary high-dimensional feature vectors without additional benefit. Within this thesis B is set to 3 since this is a reasonable compromise between descriptor size and exploitation of spatial information [50].

The ELTP descriptor is an extension of the well-known texture representation based on Local Ternary Patterns (LTP) proposed by Tan and Triggs in [40]. To recall from Section 2.2.1, Tan and Triggs [40] proposed to overcome the limitations of the generic Local Binary Patterns (LBP) descriptor [39] by replacing the binary encoded operator with a ternary code. Hence, the difference of the center pixel and its surrounding neighbors is encoded using three values $\{-1, 0, 1\}$ according to a user specified threshold t. Thus, the thresholding function of LTP $s(z_p, z_c, t)$ is defined as

$$s(z_p, z_c, t) = \begin{cases} 1, & z_p \geq z_c + t \\ 0, & |z_p - z_c| < t \\ -1, & z_p \leq z_c - t, \end{cases} \tag{4.7}$$

where z_c represents the center pixel and z_p is neighboring pixel at location p. A graphical illustration of the basic LTP operator is shown in Figure 2.5. Instead of encoding the local ternary pattern in a histogram of 3^P different bins, it is split into two separate LBP maps from which histograms are calculated and concatenated to form the final descriptor. Figure 2.6 illustrates this procedure. In this thesis, $\text{LTP}^{u2}_{(P,R)}$ operators are extracted from a (P, R) neighborhood, where P represents the number of neighboring points within a radius R around the center pixel and $u2$ indicates that uniform patterns are taken into account (see Section 2.2.1 for details). Hence, the negative and the positive part of the LTP operator are treated as uniform patterns, i.e. binary codes with more than two transitions U are assigned to one single bin while there is a separate label for every binary pattern with $U \leq 2$. Ahonen *et al.* found in [50] that 85.2% of LBP patterns in the $(8, 2)$ neighborhood are uniform for images of a human face. Consequently, using non-uniform patterns would lead to sparse histograms and hence to unnecessarily large feature vectors. Secondly, experiments in various publications suggest that in general uniform LBP/LTP operators lead to more robust and accurate descriptors [39, 50, 49]. Since Tan and Triggs [40] found that for face recognition $\text{LTP}_{(8,2)}$ operators perform best, LTP-histograms are extracted around a $(P, R) = (8, 2)$ neighborhood within the proposed framework. Moreover, instead of applying a fixed user-defined threshold t, an extension called ELTP, introduced by Liao *et al.* in [345], is applied within the proposed PRF. More specifically, t is defined by the statistics of the current LTP region and is calculated by $t = \alpha \cdot \sigma$, where σ represents the standard deviation of the image patch and $0 < \alpha \leq 1$ is a user-defined scaling factor. As suggested by Liao *et al.* in [345], α was set to 0.2 for the experiments conducted in this thesis. Once the ELTP histograms of all non-overlapping regions of each GMP have been extracted, the final feature vector is given by concatenating all extracted ELTP histograms.

Note that a number of parameters have to be taken into account for the extraction of the proposed ELGTPHS descriptor. The values of all parameters necessary for ELGTPHS feature extraction have been discussed in the previous sections and are summarized in Table B.1.

Feature Space Transformation Using Locality Preserving Projections (LPP) The output of the proposed ELGTPHS descriptor is a high dimensional feature vector which is too large to perform fast and efficient face recognition in practice. Moreover, the *"curse of dimensionality"* [347] is a well known problem in machine learning and face recognition in particular. It describes the phenomenon that as data dimensionality increases it becomes more and more difficult to extract meaningful conclusions out of the data. Beyer *et al.* showed in [348] that under very general conditions all metrics and metric-like structures will suffer from the problem that if the dimensionality of the feature space is high enough the relative separation between the data points is nearly zero. In practice, feature space transformation techniques can be applied to overcome this general difficulty of machine learning. Moreover, it is well known that supervised dimensionality reduction techniques can help classifiers to perform better in many applications since samples that belong to the same class are mapped closer together while items from different classes are mapped further apart. This fact often helps classification algorithms to better discriminate between classes. Therefore, dimensionality reduction is an important step within the proposed PRF.

Fortunately, as stated in Section 2.3, it can be shown that the intrinsic dimensionality of the data is much lower than the original feature space. Hence, the high-dimensional feature vectors $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ of size n can be projected into a smaller dimensional subspace of size m which preserves most of the discriminating information by applying a unitary projection matrix $\boldsymbol{W} \in \mathbb{R}^{n \times m}$.

$$\boldsymbol{y}_k = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{x}_k; \quad \text{with} \quad \boldsymbol{x}_k \in \mathbb{R}^{n \times 1}, \quad \boldsymbol{y}_k \in \mathbb{R}^{m \times 1}, \quad m \ll n. \tag{4.8}$$

The resulting feature vectors $\boldsymbol{y}_k \in \mathbb{R}^{m \times 1}$, with $k = 1, \cdots, N$, can then be used for classification. Two popular eigenvector-based techniques for linear feature space transformation are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Both dimensionality reduction algorithms were reviewed in Section 2.3.

One major disadvantage of both methods is that they only see the global euclidean structure of the feature space. In face recognition however it is known that "[...] the variations between the images of the same face due to illumination and lighting direction are almost always larger than image variations due to a change in face identity [...]" [349]. Although the proposed ELGTPHS descriptor is designed to be robust against those kinds of image variations, this statement is still true. A number of non-linear techniques to generate a low-dimensional representation from a high-dimensional ambient space have been proposed over the past decade [53, 54, 55, 56, 57].

However, as stated by He *et al.* in [58], although those methods achieve impressive results on artificial benchmark datasets, the obtained mappings are defined only on the training set and it often remains unclear how to apply the obtained mapping to unseen test data.

Also generalized versions of PCA and LDA were proposed which apply the so called kernel-trick for nonlinear embedding [67, 68]. However, these methods do not explicitly consider the special manifold structure of the feature space. Thus, a linear feature space transformation technique called LPP is proposed for dimensionality reduction in this thesis. LPP, developed by He *et al.* in [58], finds an embedding that preserves local information and thus fulfills the requirements for robust and efficient face recognition of great apes.

To recall from Section 2.3.3, LPP first defines an adjacency graph $\boldsymbol{\mathcal{G}}$ and an edge is put between two nodes k and j if they are "close", i.e. if they are within a user-defined ϵ-neighborhood. Thus, in the original implementation of LPP, $\boldsymbol{\mathcal{G}}$ is modeled in an unsupervised fashion. However, the optimal choice of the distance-threshold ϵ often remains unclear and highly depends on the data and the application. Hence, in this thesis the nearest-neighbor graph is modeled in a supervised fashion which is known to usually give better results than unsupervised approaches for subsequent classification [62, 350]. Therefore, an edge is put between two nodes k and j if they belong to the same class: $\boldsymbol{c}_{\boldsymbol{x}_k} = \boldsymbol{c}_{\boldsymbol{x}_j}$. Hence, the calculation of the weights from Equation 2.15 is modified as follows:

$$\boldsymbol{S}_{k,j} = \begin{cases} e^{-\frac{\|\boldsymbol{x}_k - \boldsymbol{x}_j\|^2}{2\sigma^2}}, & \text{if } \boldsymbol{c}_{\boldsymbol{x}_k} = \boldsymbol{c}_{\boldsymbol{x}_j} \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

Wang *et al.* investigated the influence of the heat kernel parameter σ on face recognition performances of the *Laplacianfaces* algorithm on various public benchmark datasets in [351]. They found that its performance is rather insensitive to the choice of σ and the experimental results suggest that setting $\sigma = 100$ is reasonable for most applications. Furthermore, the original features are first projected into the PCA subspace prior to LPP by deleting the smallest principle components which not only speeds up the performance but is also known to perform better then PCA or LPP alone [109][2]. Thus, the final embedding is as follows:

$$\boldsymbol{W} = \boldsymbol{W}_{\text{PCA}} \boldsymbol{W}_{\text{LPP}}, \tag{4.10}$$

where $\boldsymbol{W}_{\text{PCA}}$ is the transformation matrix of size $n \times m$ and $\boldsymbol{W}_{\text{LPP}}$ is the projection matrix of size $m \times m$ resulting from the subsequent LPP algorithm.

[2] Recall that this procedure was also proposed by [62] for the Fisherfaces algorithm to overcome the singularity problem of the within-class scatter matrix $\boldsymbol{S}_{\text{w}}$ (see Section 2.3)

Hence, the final transformation can be obtained by

$$\boldsymbol{x}_k \longmapsto \boldsymbol{y}_k = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{x}_k. \tag{4.11}$$

The choice of m is one of the most critical ones within the proposed PRF. Too high dimensional feature vectors compared to the sample size entail the risk of the *curse of dimensionality* while too small dimensional descriptors might produce bad results since too much information was discarded. A good choice of the dimensionality of the feature space strongly depends on the size of the dataset which is used to train the system [347].

The dataset used for evaluation of the *Randomfaces* algorithm by Wright *et al.* [71] contained 700 training images while the length of the feature vectors after random projection was set to 540. Therefore, the ratio of training images to the length of feature vectors was set to approximately 1.3 for the widely acclaimed *Randomfaces* algorithm. The smallest dataset used for experimentation in this thesis consists of 572 facial images of 24 individuals[3]. About $\frac{9}{10}$ of this data is used for training. Therefore, the size of the feature vectors after projection was set to $m = 160$ which - given the size of the utilized datasets - is small enough to overcome the *curse of dimensionality* but at the same time large enough to maintain most of the descriptor's discriminative capability. In other words, for the smallest dataset used for experimentation more than three times more samples than features were utilized to learn the task at hand. For the second dataset, even five times more samples than features were within the training database.

Classification Using Sparse Representation Classification (SRC) For the classification of global features, the SRC paradigm developed by Wright *et al.* [71] is used to identify individual great apes. A brief introduction to Compressed Sensing (CS) theory and SRC can be found in Section 2.4.2. As already discussed, the core idea of SRC is to represent a test sample $\boldsymbol{t}$ as a linear combination of the training samples $\boldsymbol{A}$

$$\boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}_1, \tag{4.12}$$

where the sparse coefficient vector $\boldsymbol{p}_1$ can be found by solving a convex optimization problem via ℓ_1-norm minimization. Motivated by CS theory, the development of fast and accurate algorithms for ℓ_1-norm minimization has received significant attention during the past years.

[3] Detailed information about all datasets used for evaluation can be found in Table 5.3.

An overview and comparison of existing algorithms can be found in [352][4]. Within the proposed PRF, the Gradient Projection for Sparse Reconstruction (GPSR) algorithm proposed by Figueiredo *et al.* in [353] is utilized for solving the ℓ^1-norm minimization problem[5].
Note that Figueiredo *et al.* solve the convex unconstrained optimization problem

$$\arg\min_{\boldsymbol{p}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{p}\|_2^2 + \tau\|\boldsymbol{p}\|_1, \tag{4.13}$$

where τ is a real-valued non-negative parameter of the objective function. It can be shown via convex analysis that problem 4.13 is closely related to the convex optimization problem given in Equation 2.28.

Experimental comparisons of GPSR to other state-of-the-art greedy scheme-based optimization algorithms such as Matching Pursuit (MP) and Orthogonal Matching Pursuit (OMP) [354, 355] showed that the proposed method is significantly faster, more efficient, and at the same time more accurate than competing algorithms. Moreover, it has been shown that GPSR works well across a large scale of applications without the requirement for application-specific tuning. For details regarding the mathematical justification and the functionality of GPSR as well as detailed comparisons to other approaches, the interested reader is referred to [353].

As suggested in [353], the parameter τ is set to $\tau = 0.1\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}\|_\infty$ in this thesis, where $\|\cdot\|_\infty$ denotes the ℓ_∞-norm which in finite dimensional space is equivalent to the supremum of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$, $\sup\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}\}$.

To recall the main idea of SRC-based classification it is summarized in algorithm 1.

As done in [71], a simple example is conducted to show the superiority of ℓ_1-norm minimization over ℓ_2-norm minimization in the context of automatic face recognition. A facial image of a chimpanzee is first downsampled to a size of 15×10 and subsequently vectorized. Thus, the 150-dimensional vector of gray-level intensities serves as feature vector $\boldsymbol{t}$. Figure 4.8 compares the results of Equation 2.28 (a) and 2.26 (c) after solving the underdetermined system of linear equations $\boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}$ using ℓ_1-norm minimization and ℓ_2-norm minimization, respectively. Furthermore, in Figure 4.8(a) two example images are plotted that correspond to the two maximum values in $\hat{\boldsymbol{p}}_1$. Additionally, Figures 4.8(b) and 4.8(d) show the residuals of Equation 2.29 for ℓ_1-norm and ℓ_2-norm minimization, respectively. For ℓ_1 minimization the test image $\boldsymbol{t}$ is correctly identified as subject 1, while for ℓ_2-norm minimization the test image is misclassified as class 12 which is due to the dense coefficient vector $\hat{\boldsymbol{p}}_2$.

[4]The "ℓ_1-magic" toolbox, a collection of MATLAB routines for solving convex optimization problems via ℓ_1-norm minimization, has recently been published and can be downloaded at `http://users.ece.gatech.edu/~justin/l1magic`. Last visit: April 2^{nd}, 2014

[5]A MATLAB implementation of GPSR can be found at `http://www.lx.it.pt/~mtf/GPSR/`. Last visit: April 2^{nd}, 2014

Algorithm 1 Sparse Representation-based Classification (SRC)

Require: $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_C] = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_l] \in \mathbb{R}^{m \times l}$: transformed matrix of training samples
$\boldsymbol{t} \in \mathbb{R}^{m \times 1}$: transformed test sample
$\boldsymbol{\delta}_i \in \mathbb{B}^{l \times 1}$ for $i = 1, \cdots, C$: characteristic function for every class

function SRC($\boldsymbol{A}, \boldsymbol{t}, \boldsymbol{\delta}_i$)

  for $k \leftarrow 1$ to l **do**
    $\boldsymbol{a}_k \leftarrow \frac{\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|}$ ▷ Normalize columns of $\boldsymbol{A}$ to have unit ℓ^2-norm
  end for
  $\boldsymbol{t} \leftarrow \frac{\boldsymbol{t}}{\|\boldsymbol{t}\|}$ ▷ Normalize $\boldsymbol{t}$ to have unit ℓ^2-norm
  $\hat{\boldsymbol{p}}_1 \leftarrow \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\|_1$ subject to $\boldsymbol{t} = \boldsymbol{A}\boldsymbol{p}$ ▷ Solve the convex optimization problem via ℓ^1-norm minimization
  for $i \leftarrow 1$ to C **do**
    $r_i(\boldsymbol{t}) \leftarrow \|\boldsymbol{t} - \boldsymbol{A}(\boldsymbol{\delta}_i \odot \hat{\boldsymbol{p}}_1)\|_2$ ▷ Compute residuals for every class
  end for
  $\mathrm{ID}(\boldsymbol{t}) \leftarrow \arg\min_i r_i(\boldsymbol{t})$ ▷ The identity is the class with the minimum residual
  return $\mathrm{ID}(\boldsymbol{t})$

end function

As can be seen from Section 4.2.3.1 and according subsections, a number of parameters for global feature extraction, feature space transformation, and classification exist. A thorough tuning and experimental evaluation of all these parameters across their entire ranges is out of the scope of this thesis and is even infeasible in practice due to limited data sizes. Hence, all parameters were set to established and reasonable default values as described above. Table B.1 gives an overview of the applied parameters for the global face recognition pipeline.

4.2.3.2. Face Recognition Using Local Features

It is well known from psychophysics and neuroscience that both holistic and local information are important for perception and recognition of faces [327, 328, 329]. Additionally, it has been reported in the literature that different feature representations tend to misclassify different patterns [356]. Therefore, global and local features offer complementary information which can be used to improve robustness and accuracy of the proposed system. While for global face representation ELGTPHS features are proposed (see Section 4.2.3.1), Speeded-Up Robust Features (SURF) (see Section 2.2.2) around certain facial keypoints are used as local descriptors. While global descriptors represent the whole appearance of an ape's face, local features should be more robust against local changes as they only encode detailed traits of the corresponding point of interest. More precisely, distinctive wrinkle patterns around eyes and nose might offer individually discriminative information which is implicitly encoded by SURF. The subsequent paragraphs explain the face recognition pipeline using local features in more detail.

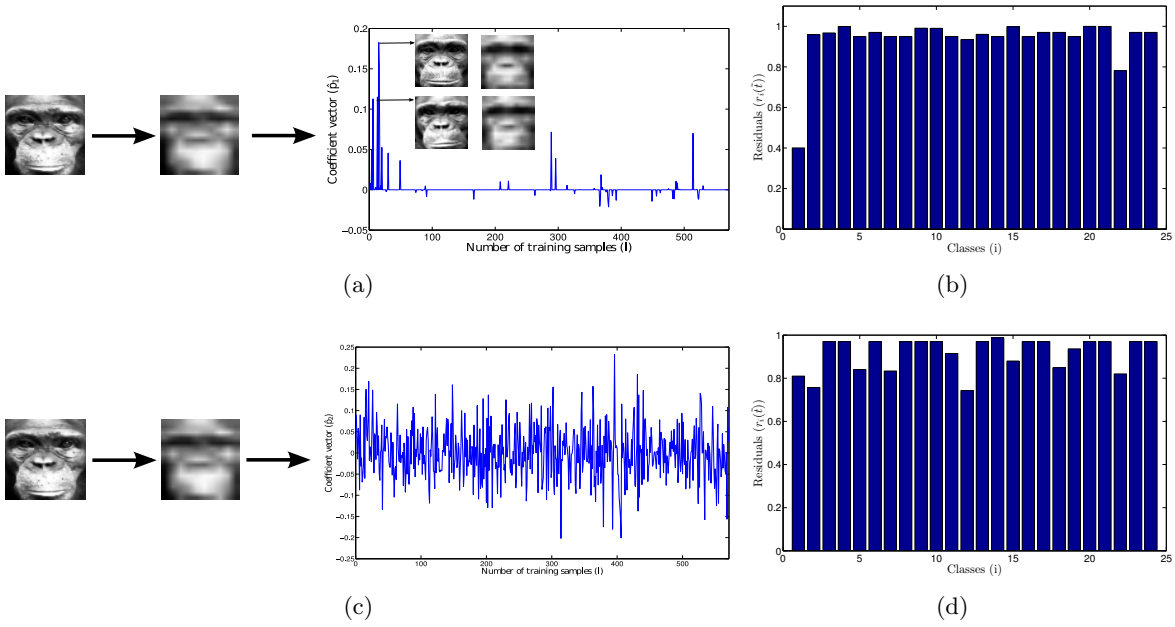


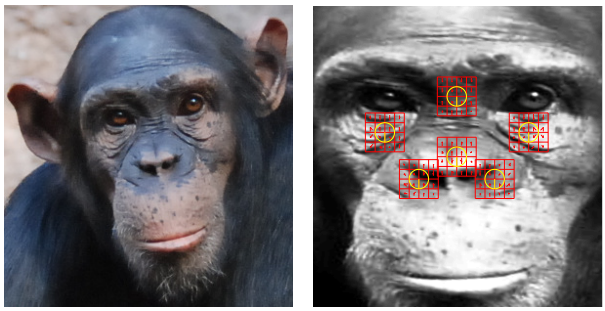
Figure 4.8.: ***Comparison of classification using ℓ_1-norm and ℓ_2-norm minimization.*** A test image is downsampled to a size of 15×10 and subsequently vectorized to produce the feature vector $\boldsymbol{t}$. It is then classified by SRC. The sparse vector $\hat{\boldsymbol{p}}_1$ and two example images corresponding to the two largest entries in $\hat{\boldsymbol{p}}_1$ are plotted in (a) after solving the ℓ_1-norm minimization problem (see Equation 2.28). The according residuals are shown in (b). The test image is correctly identified as subject 1 (smallest residual). The results are compared with the vector $\hat{\boldsymbol{p}}_2$ (c) and the according residuals (d). It it obvious that the face recognition problem cannot be solved correctly by ℓ_2-norm minimization due to the dense property of $\hat{p}_2$.

Local Keypoint Description using Speeded-Up Robust Features (SURF) As stated in Section 2.2.2, SURF is an efficient scale- and rotation-invariant local keypoint detector and descriptor first proposed by Bay *et al.* in [52]. Since the keypoint locations are calculated based on the eye and mouth positions obtained from the face and facial feature detection library SHORE™ (see Section 4.2.1), the detector part of SURF can be omitted for the application presented in this thesis. Furthermore, interest points are usually detected at different scales - in the so called scale-space - in order to achieve scale invariance of the descriptor. However, after face alignment (see Section 4.2.2) the size of the resulting face image is known and thus the regions of interest around all facial keypoints can be derived based on height and width of the aligned facial image. Hence, instead of estimating the scale of each interest point it can be set automatically, i.e. SURF features can be calculated only in one particular scale which is constant for all keypoints of a given face. Another key step of SURF is to identify the dominant orientation of the interest point in order to obtain rotation invariance of the descriptor. This step can also be omitted for the proposed application since face images are already in an upright position after

alignment. According to [52], the upright version of SURF (hereafter denoted as U-SURF) is faster to compute and can even increase distinctiveness while maintaining a certain robustness against small rotation angles of up to $\pm 15°$. Omitting all these pre-processing steps of the classic SURF descriptor significantly speeds up the performance of the system without suffering loss of robustness which facilitates near real-time performance of the system.

For the proposed application, U-SURF descriptors are extracted around 6 facial keypoints which are calculated with regard to the detected eye and mouth markings obtained from SHORE™ (see Section 4.2.1). Based on the assumption that wrinkle patterns under and between the eyes are unique across individuals and useful for identification, the first three points are located under the left and right eye, as well as between both eyes. Furthermore, it is assumed that the area around the nose is well suited for discrimination. Therefore, the tip of the nose as well as the left and the right nostril serve as additional locations for local feature extraction. The mouth region is not used for local feature extraction because this area is often occluded and deformed due to eating and facial expressions. Extracting features in this region might lead to a high intra-class variance and would therefore hamper classification. Note that in order to adequately encode the gradients of wrinkle patterns all facial images are only aligned and converted to gray-scale without any scaling. However, as done for the proposed global face recognition pipeline the minimal images size is again set to 64×64 pixels. Thus, SURF features are extracted within the originally sized face images in order to capture as much individual specific information as possible. The width of the region of interest around each keypoint was empirically set to be approximately $\frac{1}{7}$ the width of the original aligned facial image which was found to suitably overlap with the according wrinkle patterned region. Figure 4.9(b) shows the location of the facial markings which were used for local feature extraction. The extracted Haar wavelet responses as well as the rectangular square region for every interest point are superimposed. All six 64-dimensional local SURF descriptors are subsequently concatenated to form one single representation for each face resulting in a feature vector $\boldsymbol{x}_{\mathrm{SURF}}$ of size $6 \cdot 64 = 384$.

Feature Space Transformation of Local Features As done for the global face descriptor, LPP is used to transform the concatenated local feature vector into a smaller dimensional subspace to speed up the subsequent classification and avoid the *curse of dimensionality*. Again a heat-kernel is used to model the symmetric weight-matrix $\boldsymbol{S}$ in a supervised fashion (see Equation 4.9). Furthermore, PCA is applied in order to first delete the smallest principle components before LPP (see Equation 4.10). All parameters are set as discussed for the global face recognition pipeline in Section 4.2.3.1.



(a) (b)

Figure 4.9.: ***Aligned chimpanzee face and superimposed keypoints for U-SURF extraction.*** (a) A detected chimpanzee face. (b) The aligned face and the superimposed locations of the keypoints. The U-SURF descriptors are extracted around six facial keypoints which were calculated based on the eye and mouth markers obtained from SHORE™ in order to utilize the uniqueness of the wrinkle patterns. A grid comprising 4×4 sub-regions (illustrated in red) is created around every landmark point from where the gradient directions and magnitudes are calculated to build the final U-SURF descriptor. The yellow circles denote the scale and the main orientation. Note that both parameters are set to fixed values in this application since the face images are already aligned and the size of the resulting facial image is known.

Classification using Support Vector Machines(SVMs) The proposed system uses an SVM for the classification of local features. As introduced in Section 2.4.3, an SVM is a discriminative classifier, attempting to generate an optimal decision plane between feature vectors of the training classes. In many real-world applications, classification with linear separation planes is often not possible in the original feature space. By utilizing a so-called kernel trick, the feature vectors can be efficiently transformed into a higher dimensional space in which they are linearly separable. In this thesis, an Radial Basis Function (RBF) kernel is applied to transform the data into a higher dimensional space:

$$\mathcal{K}(\boldsymbol{y}_i, \boldsymbol{y}_j) = e^{-\gamma \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}, \tag{4.14}$$

It can be seen from Equations 2.31 and 4.14 that there are two parameters (C, γ) that need to be optimized during training. As recommended by Hsu *et al.* in [78], a grid search is applied for this purpose by using a 5-fold cross-validation on the training set. Exponentially growing sequences of (C, γ) value pairs are evaluated and the one with the best accuracy is applied to build the final model. Hsu *et al.* suggest to search the following sequences of C and γ for an appropriate parameter combination in practical applications: $C = 2^{-5} \ldots 2^{15}$ and $\gamma = 2^{-15} \ldots 2^{-3}$. Since testing every combination of C and γ can take a significant amount of time, a coarse-to-fine grid search is applied. After the best combination of C and γ is found on a coarse grid, this region is subsequently scanned more thoroughly on a finer resolved grid.

Hsu *et al.* recommend to repeat this procedure three times with decreasing step sizes for C and γ until the best parameter setting was found. The applied parameter ranges and step sizes were applied as suggested in [78] and are summarized in Table C.1.

As already mentioned in Section 2.4.3, SVM in its original implementation is a binary classifier which can only handle two-class problems. In this thesis a *one-vs-one* approach [357] is used to handle the multi-class problem of face recognition. Multiple binary classifiers are built which are able to distinguish between every pair of classes. Classification can then be done by a simple voting strategy, in which every binary classifier assigns the test sample to one of the two classes which is consequently increased by one. Finally, the class with the most votes determines to which class the test image belongs. The freely available Library for Support Vector Machines (LIBSVM) toolbox [77], a software library for support vector classification, is utilized in this thesis for training and testing. By default, SVMs only predict the class a test vector $\boldsymbol{t}$ belongs to. However, Chang *et al.* extended LIBSVM to give probability estimates based on ideas of [358]. This section briefly describes the LIBSVM implementation of probability estimation in SVMs. For details and mathematical justification of the following equations, the interested reader is referred to the publications by Wu *et al.* [358] and Chang *et al.* [77]. The probability estimates obtained by *pairwise coupling* are used as confidence measures for classification of local features. Given C classes, for every test sample $\boldsymbol{t}$ the objective is to estimate

$$p_i = P(\boldsymbol{c}(\boldsymbol{t}) = i|\boldsymbol{t}), \quad i = 1, \ldots, C, \tag{4.15}$$

where $\boldsymbol{c}(\boldsymbol{t})$ is the class prediction of test sample $\boldsymbol{t}$. Following the one-vs-one approach for multi-class classification of SVM, *Platt Scaling* [359] is used to estimate the pairwise class probabilities by using a logistic sigmoid-function

$$r_{ij} = P(\boldsymbol{c}(\boldsymbol{t}) = i|\boldsymbol{c}(\boldsymbol{t}) = i \text{ or } j, \boldsymbol{t}) \approx \frac{1}{1 + e^{A\hat{f}+B}}, \tag{4.16}$$

where $\hat{f}$ is the decision function of SVM and A and B are parameters that are estimated during training. After collecting all r_{ij} values, Wu *et al.* [358] propose to solve the following optimization problem in order to obtain the class probabilities for each class:

$$\min_p \frac{1}{2} \sum_{i=1}^{C} \sum_{j:j\neq i} (r_{ji}p_i - r_{ij}p_j)^2 \qquad \text{subject to} \qquad \sum_{i=1}^{C} p_i = 1, \quad p_i > 0 \quad \forall i \tag{4.17}$$

Again, a number of parameters have to be set for the local face recognition pipeline. The settings of all parameters were discussed within the previous sections. Table C.1 gives an overview of all parameters for face recognition using local features.

4.2.3.3. Decision Fusion

After classifying global and local descriptors separately, the final objective is to fuse the results of both recognition pipelines in order to obtain a single final decision. Based on the observation that different features tend to misclassify different patterns [356], both face identification paradigms should benefit from each other. It has been demonstrated in many publications that combining information from multiple sources can enhance a system's robustness and therefore improve its accuracy. In general, two different fusion methodologies can be distinguished: *feature-level fusion* and *decision-level fusion*. As the name suggests, feature-level fusion strategies try to combine descriptors from different sources in the actual feature space before classification. Typical representatives ranging from simple feature vector concatenation techniques to more advanced techniques such as Canonical Correlation Analysis (CCA) [360, 361] or Kernel Canonical Correlation Analysis (KCCA) [362]. Decision-level fusion techniques on the other hand combine the classification results of different modalities to enhance the system's performance. In this thesis the latter methodology is applied since studies suggest that decision-level fusion techniques are superior to feature-level fusion methods for many applications [363, 364]. A comprehensive review of decision-level fusion strategies can be found in [365]. According to [365], three types of decision-level fusion methods can be differentiated. Techniques which operate only based on the predicted class labels utilize the minimum amount of information, while additional information might be gained from class rankings. However, fusion methods which operate on fuzzy outputs, i.e. the confidences of multiple classifiers, can be expected to produce the greatest improvement. Therefore, a simple but effective decision fusion paradigm is proposed in this section which not only takes the class rankings of each classifier into account but additionally utilizes the confidences of both classifiers to weight their predictions accordingly.

Figure 4.5 illustrates the parallel fusion scheme that is applied in this thesis. The decision-level fusion paradigm of the proposed PRF was influenced by ideas of [366]. A parallel fusion strategy which combines the rank-outputs of different classifiers is used to fuse the results of local and global features. The parallel fusion scheme proposed in [366] only uses a single weighting function $\boldsymbol{w}(\wp) = \wp^c$ for rank $\wp = \{1, \cdots, C\}$ and constant c to combine the outputs of two classifiers. In this thesis, a non-linear rank-sum method is proposed which weights the results of both classifiers using different weighting functions. Additionally, the confidences of each classifier are taken into account when generating the weighting function $\boldsymbol{w}(\wp) = e^{\boldsymbol{s}(\wp)}$, where $\boldsymbol{s}(\wp)$ represents the confidence of SRC and SVM for rank $\wp$, respectively. For SRC, the vector of residuals from Equation 2.29 is used as confidence measure, while for SVM the probability estimates of LIBSVM [77, 358] can be utilized. Note, however, that for SRC the *minimal* residual determines the class affiliation, while for SVM the test sample is assigned to the class with the *maximum* probability. Fortunately, the residuals of SRC can simply be converted into confidence measures

by negating the residuals. Thus, the score vector of SRC is given by $\boldsymbol{s}_{\mathrm{SRC}} = 1 - \boldsymbol{r}$, where $\boldsymbol{r}$ is the vector of residuals from Equation 2.29. The final score vector $\boldsymbol{s}_{\mathrm{df}} \in \mathbb{R}^{C\times 1}$ is then simply given by the sum of both weighting functions: $\boldsymbol{s}_{\mathrm{df}} = \boldsymbol{w}_{\mathrm{SRC}} + \boldsymbol{w}_{\mathrm{SVM}}$, where C is the number of classes and the weight vectors are defined as

$$\boldsymbol{w}_{\mathrm{SRC}} = e^{\boldsymbol{s}_{\mathrm{SRC}}} = e^{1-\boldsymbol{r}} \qquad \text{and} \tag{4.18}$$

$$\boldsymbol{w}_{\mathrm{SVM}} = e^{\boldsymbol{s}_{\mathrm{SVM}}}. \tag{4.19}$$

Finally, $\boldsymbol{s}_{\mathrm{df}}$ is sorted in descending order to obtain the final result which is the first entry of the sorted score vector.

4.3. Face Recognition in Video

After explaining the core algorithms of the proposed facial recognition system for great apes, this section extends these ideas to perform robust and accurate identification in video sequences. Figure 4.10 gives an overview of the extended PRF. Components that are also used for face recognition in images are marked with a blue frame while new or significantly modified processing steps are illustrated in red.

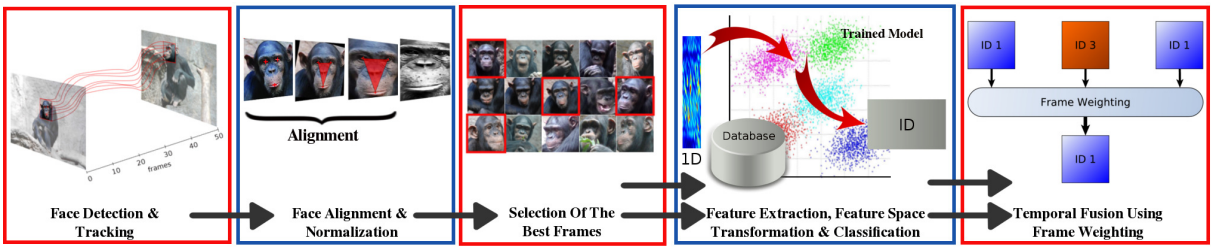


Figure 4.10.: ***Overview of the proposed system for primate face identification in videos.*** New processing steps are marked with a red frame while components that have been used for face recognition in still images are highlighted in blue. One key-step for face recognition in video is to simultaneously track multiple faces. Once each tracked target is assigned to a face-track, faces are first aligned as done for still images. Subsequently, modules for best-frame-selection are applied which analyze parameters such as pose and visual quality. Each selected and aligned face is subsequently identified by means of the feature extraction, feature space transformation, and classification algorithms explained in the previous sections. A frame-weighting approach is proposed which implicitly exhibits the temporal information of video recordings to further enhance the system's accuracy.

A five step approach to perform robust and accurate identification of great apes in video sequences is proposed and discussed in detail within the subsequent sections:

1. In addition to the detection of faces in still images, tracking of faces is a prerequisite for identification of great apes in videos. Therefore, unique object-IDs are assigned to each detected face which are maintained for the subsequent frames. This procedure results in a so called face-track, a collection of faces from one single individual in various appearances.

2. For every face within each face-track the same alignment and normalization techniques discussed in Section 4.2.2 are applied in a subsequent processing step.

3. Usually not all face images of an obtained face-track are well suited for face recognition. Hence, in a third step the faces that are best suited for recognition in terms of head pose and visual quality are automatically selected and subsequently transferred to the main identification modules.

4. In a fourth step, local and global features are extracted from each normalized face, transformed into a smaller dimensional subspace, and classified using SRC and SVM as explained in Section 4.2.3.

5. Finally, the obtained results from the identification step are combined by means of a novel frame-weighting approach which implicitly utilizes the temporal information contained in videos. The results of multiple automatically selected frames are weighted and combined by taking the classifier's confidence of each frame into account.

4.3.1. Face Tracking by Continuous Detection

One of the key steps to perform facial identification of great apes in video footage is to track detected faces through the sequence. As stated in Section 4.2.1, SHORE™ is not only capable of detecting faces in single frames but also track them through a scene. Once a face has been detected, a unique ID is assigned to the detected object. During consecutive frames, the tracking algorithm then tries to maintain the association between object-ID and the detected face. This section briefly reviews the tracking algorithm used within SHORE™. For a more detailed explanation the interested reader is referred to [343].

Küblbeck and Ernst suggested a tracking by continuous detection approach in [343] to overcome the deficiencies of pure tracking algorithms. Each frame is processed with a face detector as described in Section 4.2.1. A temporal scanning procedure called *slicing* is utilized to ensure real-time capability of the system. The static detector repeatedly searches for faces in all levels of an image pyramid in order to find faces of different sizes. The main assumption of slicing is that in most practical applications it is very unlikely that a face changes its size drastically in a fraction of a second. Hence, it is sufficient to scan all pyramid levels only a few times per second instead of applying face detection for all resolutions in each frame. Therefore, the image

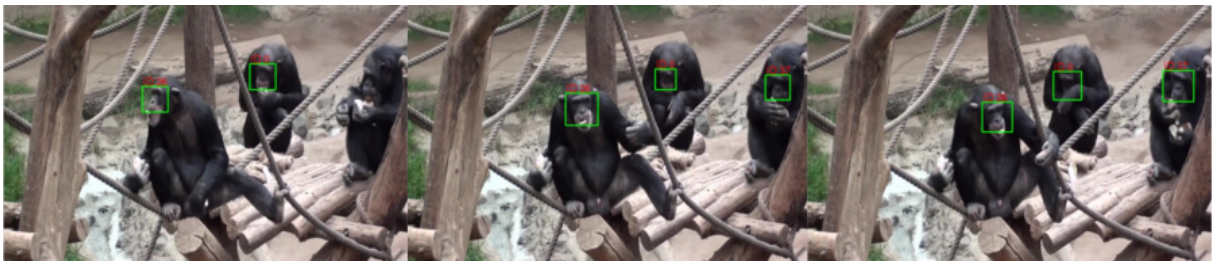
pyramid is partitioned into slices of almost equal size which are then processed sequentially, i.e. with every new frame another part of the image pyramid is scanned for faces. For instance if the image pyramid is partitioned into five equally sized parts, each slice is processed every fifth frame which theoretically speeds up the performance of the face detection system by a factor of five. In practice, however, detection is performed not only in a single pyramid level but in neighboring levels in order to ensure that tracked faces get updated in every frame which eventually leads to a smaller speedup than theoretically possible. In practical applications Küblbeck and Ernst observed a performance improvement by a factor of two to three, depending on the number of faces in the scene [343]. A motion model is then applied to connect the detections of subsequent frames. A linear Kalman filter [137, 138] is applied in order to estimate the current state of a tracked face from the detection results. An object state comprises the spatial position as well as the distance between both eyes which is an estimate of the size of a detected face. Additionally, the first and second order derivatives are included in the state vector to represent the velocity and the acceleration of a face. Association of object-ID and detected face in consecutive frames is done by using a minimum distance criterion: A detected face in the current frame is associated with the face detected in the previous frame which is closest to the current object position. Furthermore, it was shown in [343] that based on the observations of the past frames it can be decided if a tracked object actually represents a valid face which significantly reduces the number of false positive detections while the detection rate is maintained.

Figure 4.11 shows an excerpt of a video sequence gathered in the zoo of Leipzig and an extracted face-track for one of the individuals present in the video.

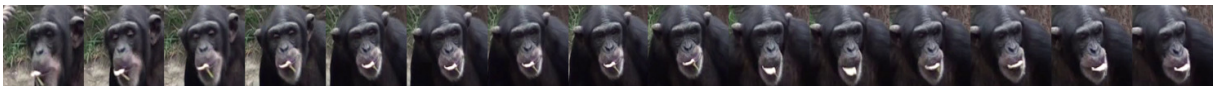
Three individuals are present in the scene and the faces are automatically detected and tracked by SHORE™. The detection results and associated object-IDs are superimposed in Figure 4.11(a). A face-track, a collection of faces detected in consecutive frames from a single individual in various appearances, can then be generated for each target. A few frames of an extracted face-track are illustrated in Figure 4.11(b). Since the proposed PRF is optimized for full-frontal faces, not all frames of a face-track are equally well suited for automatic identification. Hence, the frames which are best suited for recognition are automatically selected in a second step of the proposed system which involves multiple parameters such as pose and visual quality.

4.3.2. Selection of the Best Frames

This section describes algorithms for full-automatic selection of those frames of a face-track which are best suited for identification. The most important cues to achieve high recognition accuracies in practical applications are the pose of a face and its visual quality. The automatic analysis of both parameters is discussed below.



(a)



(b)

Figure 4.11.: ***Chimpanzee faces tracked by SHORE™.*** (a) Three selected frames of a video gathered in the zoo of Leipzig, Germany. The detection and tracking results by SHORE™ are superimposed. The face of the chimpanzee in the front moves significantly through the scene. The tracking algorithm assigns an unique object-ID to each detected face and tries to maintain the given ID until an individual leaves the scene or cannot be tracked anymore due to occlusion or missed detections. (b) An extracted face-track of one of the three individuals in the video sequence. Obviously not all frames of a face-track are equally well suited for recognition. [I04]

4.3.2.1. Pose Estimation

Automatic head pose estimation has been an active research topic for decades. According to [367], head pose estimation is the process of automatically estimating the orientation of a (human) head. This procedure requires a number of processing steps to transform a pixel-based representation into a high-level concept of direction relative to a global coordinate system. As humans and great apes share most of their facial appearance properties, algorithms commonly used for human head pose estimation should therefore be applicable to great apes as well.

The head of humans and great apes is limited to three degrees of freedom which are commonly referred to as *pitch*, *roll*, and *yaw* angles. A variety of state-of-the-art algorithms were proposed over the last decades ranging from coarse-level approaches that simply differentiate between few discrete directions (e.g. frontal versus profile poses) to more granular techniques that try to estimate a continuous angular measurement of the head pose in multiple degrees of freedom. A comprehensive state-of-the-art literature review on head pose estimation techniques is out of the scope of this thesis and the interested reader is thus referred to the survey of Murphy-Chutorian and Trivedi [367] for instance.

For the application presented in this thesis, however, an exact estimation of pitch, roll, and yaw angles in 3D space is not required. Instead, the objective of the proposed pose estimation module is to automatically select those frames within a face-track that contain faces in a near-frontal pose. Hence, the problem of pose estimation eventually breaks down to a two-class classification problem: *frontal* versus *non-frontal* faces. Thus, a lightweight pose estimation module is proposed with regard to the following requirements:

- **Low Resolution:** Many state-of-the-art algorithms apply sophisticated flexible models such as Active Appearance Models(AAMs) or Active Shape Models(ASMs) in order to accurately estimate the pose of the head. These techniques often require high-resolution data to perform well. However, surveillance cameras which are commonly used by biologists to gather video material from great apes often do not record such high-resolution video data. Thus, the proposed algorithm should perform equally well for both high and low resolution data.

- **Illumination Variation:** The head pose estimation algorithm should be robust to the challenging lighting conditions often present in real-world environments.

- **Performance:** Unlike the actual identification of great apes which only needs to be done in a few selected frames, head-pose estimation needs to be done in each frame of each face-track in a video sequence. Thus, compact yet discriminating features and fast classification are required to achieve near real-time capability of the overall face recognition system.

In order to meet the above mentioned design criteria, the following head pose estimation technique is proposed which performs sufficiently well for the practical real-world application presented in this thesis. As illustrated in Figure 4.10, all detected or tracked faces first have to be aligned. To speed up the performance of the system, the same face alignment procedure as proposed in Section 4.2.2 is used which has to be applied for accurate recognition anyway. Each face image is subsequently converted to gray-scale, scaled to a size of 64×64 pixels, and Gabor-based features are extracted. An introduction of Gabor kernels has been given in Section 2.2.1. Each aligned face image is initially convolved with multi-scale and multi-orientation Gabor kernels (see Equation 2.2). However, in contrast to the face recognition algorithm presented in Section 4.2.3.1, a faster but more lightweight descriptor is used for pose estimation. Thus, Gabor Magnitude Picture (GMP) are used directly as features which can not only be efficiently computed but are also discriminating enough to distinguish between frontal and non-frontal faces. Moreover, the resulting GMPs can be reused for extracting the ELGTPHS descriptor proposed in Section 4.2.3.1 which builds the basis for primate identification.

Starting from the set $\boldsymbol{\mathcal{S}} = \{|\boldsymbol{G}_{\mu,\nu}(z)| : \nu \in \{0,\ldots,S\}, \mu \in \{0,\ldots,F\}\}$ which forms the overall GMP representation of the aligned face image, the overall feature vector is constructed as

$$\boldsymbol{x}_{\text{GABOR}} = \left(\boldsymbol{g}_{0,0}^{(\rho)}, \boldsymbol{g}_{0,1}^{(\rho)}, \cdots, \boldsymbol{g}_{1,0}^{(\rho)}, \cdots, \boldsymbol{g}_{S,F}^{(\rho)}\right), \tag{4.20}$$

where $\boldsymbol{g}_{\nu,\mu}^{(\rho)}$ is a column vector representing the normalized and vectorized version of the magnitude matrix $|\boldsymbol{G}_{\mu,\nu}(z)|$ which is downsampled by factor ρ.

The generation of the Gabor kernels is done as described in Section 4.2.3.1. All parameters remained as discussed above and are listed in Table B.1. After convolving the facial image with the resulting 40 Gabor wavelets, the magnitude matrices are downsampled by a factor of $\rho = 8$ using a bilinear interpolation. As suggested by Yang *et al.* in [48] a downsampling factor of 8 is a good trade-off between descriptor size, compactness, and discriminating capabilities for most real-world applications.

A PCA is applied (see Section 2.3.1) to project the resulting high-dimensional feature vectors into a smaller m-dimensional subspace. The goal of PCA is to find a subspace whose basis vectors correspond to the maximum-variance directions of the original feature space. Although more sophisticated feature space transformation techniques exist (see Section 2.3), PCA - as an example of unsupervised dimensionality reduction methods - is not only fast and easy to implement but is also expected to produce sufficiently good results for the two class problem of pose estimation. It was empirically found that a reduced feature dimension of $m = 100$ is small enough to circumvent the curse of dimensionality but also large enough to obtain good results for the task at hand [368].

An SVM with RBF kernel was used for classification. As suggested in Section 4.2.3.2, the two user-defined parameters (C, γ) were optimized during training using a grid-search in combination with a 5-fold cross-validation on the training set. All parameter ranges and step sizes for estimation of C and γ were set as recommended by Hsu *et al.* in [78] (see Table C.1). Additionally, a bootstrapping-like procedure was implemented in order to optimize the discrimination capabilities of the proposed pose estimation module: First, manually annotated face images of primate faces were divided into frontal and non-frontal faces to generate an initial model. Subsequently, this model was applied on a validation set of video data which were neither used for training nor testing during the experiments conducted in this thesis. Faces which were detected and then classified by the proposed pose estimation module were saved, manually re-annotated, and added to the training set. More specifically, faces which were falsely classified as *non-frontal* were added to the positive class. On the other hand, primate face images that were categorized as *frontal* but actually showed faces in a semi-profile or even profile pose were manually added to the negative class. Moreover, false positive detections by SHORE™ were appended to the

negative class in order to identify non-face data more easily during the subsequent processing steps. Finally, a new SVM model was then trained on the extended training database. This procedure was repeated multiple times until the resulting pose estimation model was tuned to produce adequate results on the above mentioned validation set.

The resulting model was finally saved and applied in the test case. The probability estimate for a frontal classification by the SVM prediction is utilized as a weighting factor for every face within a face-track. Hence, the *Pose-Quality* measure Q_p of the ***Pose-Module*** is a real number between 0 and 1. The higher the value, the more frontal is the pose of the considered face.

4.3.2.2. Frame Quality Assessment

The head pose is an important cue to automatically select the frames which are best suited for identification. However, a number of other parameters representing the visual quality of a face can additionally influence the results of the subsequent recognition task. In this section, a number of lightweight software modules are proposed and applied in order to estimate the visual quality of faces within a face-track.

Image Blur The ELGTPHS feature proposed in Section 4.2.3.1 as well as the SURF descriptors used for local feature extraction are compact representations of the edge and gradient information within primate faces. Thus, intuitively the sharpness of an image is an important factor to assess the visual quality of a face. The low-pass filtering characteristic of a blurred image might not contain sufficient details which are important for accurate identification.

Two important types of blur can be identified for the application in this thesis: *Out-of-focus blur* and *motion blur*. Both categories are relevant for accurate identification and should therefore be detected by the ***Blur-Module***. Although both types of blur are caused by different reasons, they result in similar image errors and can thus be reliably detected by the same algorithm [369]. The applied blur estimation module is part of *Photo Summary*[6], a technology developed by Fraunhofer IDMT through which a representative selection from a digital photo collection can be automatically compiled. This software module is influenced by ideas of Liu *et al.* published in [369]. The detection of blurred images is done by analyzing its *Local Power Spectrum Slope*.

High-frequency components in blurred images are lost due to their low-pass filtering characteristic. Hence, a number of studies demonstrated that the slope α of the power spectrum $\boldsymbol{S}(u, v)$ of a blurred image tends to be steeper than that of a sharp image [370, 371, 372].

[6] `http://www.idmt.fraunhofer.de/en/Service_Offerings/products_and_technologies/m_p/photosummary.html` Last visit: August 8^{th}, 2014

Let the power spectrum of an intensity image $\boldsymbol{I}(x, y) \in \mathbb{R}^{h \times w}$ be denoted as

$$\boldsymbol{S}(u, v) = \frac{1}{h \cdot w} |\mathcal{F}\{\boldsymbol{I}(x, y)\}|, \tag{4.21}$$

where $\mathcal{F}\{I(x, y)\}$ is the Fourier transformed image. The power spectrum is then converted into polar coordinates with $u = f\cos(\Theta)$ and $v = f\sin(\Theta)$ resulting in $\boldsymbol{S}(f, \Theta)$, where f denotes the frequency and Θ the angle. By summing the power spectra over all directions it was shown in [371] that the function $\boldsymbol{S}(f)$ is approximately given by

$$\boldsymbol{S}(f) = \sum_{\Theta} \boldsymbol{S}(f, \Theta) \approx \frac{A}{f^{-\alpha}}, \tag{4.22}$$

where A is an amplitude scaling factor and α is the slope of the power spectrum. It has been shown in large-scale experiments that for most natural images $\alpha \approx 2$. For blurred images, however, α is often larger which corresponds to a steeper slope of $\boldsymbol{S}(f)$.

Since it is hard to determine the blurriness of an image solely based on a global measurement of α, Liu *et al.* propose to utilize a relative global-to-local slope measure. In this thesis a similar approach is applied. First, the power spectrum's slope of the global image α_g is calculated. Secondly, the input image is divided into 3×3 blocks and the slope of the power spectrum is calculated separately for each block. The final metric for each patch p is then given by

$$\eta_p = \frac{\alpha_g - \alpha_p}{\alpha_g}, \tag{4.23}$$

where α_p is the slope of the power spectrum of patch p. Hence, the smaller η_p the more blurred is patch p. The overall blur measure of the facial image is finally given by averaging the η_p's of all patches. After normalization, the output of the ***Blur-Module*** is a *Blur-Quality* measure Q_b in the range $[0, 1]$. Small values correspond to blurred images while textured and thus sharp face images are represented by higher values for Q_b.

Illumination Illumination is another important indicator for visual quality which can influence automatic identification of primates positively or negatively. Hard shadows on a face for instance can often occlude important image features and consequently hamper performance of a recognition system. Similarly, overexposure and underexposure are important photometric factors which should be taken into account for an automatic face recognition application which operates in real-world environments. Although sophisticated illumination normalization techniques are capable to compensate image variability caused by extreme lighting conditions to a certain extent, it is desirable to select a shot with suitable illumination prior to identification.

A large and growing amount of literature investigates methods to automatically classify the exposure of photos based on sophisticated image features and machine learning techniques for automated organization and browsing of photo collections [12, 373, 374]. Although those approaches are known to perform well on official benchmark datasets, they appear to be too complex for an application as presented in this thesis. They often require a significant amount of computational power which consequently might hamper near real-time performance of the overall face recognition system. Therefore, the following approaches represent lightweight methods to estimate the exposure of a facial image.

First, an ***Underexposure-Module*** was implemented which estimates the degree of underexposure of a facial image. A high number of nearly black pixels suggest that an image is underexposed. First, a gray-scale histogram is created to measure the number of dark pixels in an image. The number of bins was chosen to be 20, which - considering the naturally dark pigmentation of primate faces - was empirically found to be a reasonable choice to detect underexposed face images of great apes. Hence, for a gray-scale image with a bit depth of 8 bit, i.e. 256 gray level values, the first bin of the histogram represents pixel values in a range from 0 to 12. The resulting *Underexposure-Quality* measure Q_{ue} is therefore set to 1 if there are no values in the first bin of the histogram and it is linearly decreasing with the number of pixel intensities falling into the first bin of the histogram. A completely black image, i.e. all pixel values of the image fall into the first bin of the histogram, would hence have a *Underexposure-Quality* of 0.

A second measure used within the proposed PRF is the degree of overexposure of a facial image determined by an ***Overexposure-Module***. Similar to underexposure, an overexposed image often does not contain enough information required for an accurate identification. A simple yet effective way to measure the degree of overexposure is the mean-intensity of an image $\boldsymbol{I}(x, y) \in \mathbb{R}^{h \times w}$ which is given by

$$\mu_{\mathbf{I}} = \frac{1}{h \cdot w} \sum_{x=1}^{w} \sum_{y=1}^{h} \boldsymbol{I}(x, y). \tag{4.24}$$

If $\mu_{\mathbf{I}}$ is above a certain user-defined threshold, the facial image is assumed to be overexposed. According to [375], in most consumer cameras the autoexposure system sets all parameters so that the average brightness of an image is *"middle gray"* regardless of the scene's actual brightness. Typical middle gray scenes include scenarios on overcast days under diffused light. Therefore, portrait photos taken under a cloudy bright sky are usually perfectly exposed using autoexposure [375] and are thus well suited for automatic face recognition. Therefore, the threshold for the proposed overexposure module is set to 50% of the maximal possible intensity of a gray-scale image $I_{\max}$. Thus, the *Overexposure-Quality* Q_{oe} is set to 1 if $\mu_{\mathbf{I}} \leq 50\% \cdot I_{\max}$.

If the mean-intensity of the input image is above the defined threshold, the *Overexposure-Quality* decreases linearly. Hence, a completely white image would get an *Overexposure-Quality* score of 0.

The third and final measure used to estimate the illumination quality of a face is the contrast of the input image which is estimated by a ***Contrast-Module***. Contrast is a measure for the brightness range within a scene. Hence, an image with low contrast might not contain enough information which can be utilized for identification. On the other hand, facial images with too extreme contrast ranges might also be hard to classify correctly. Since the identification of primate faces is based on gradients and pixel differences as explained in Section 4.2.3, various illuminations of different face areas might lead to false edges which could be critical for accurate classification. Furthermore, exceptionally high contrast ranges of facial images might be due to occlusion by objects with a different specular reflectance than the actual face. The Root Mean Square (RMS) contrast which measures the standard deviation of the pixel intensities is used to assign a *Contrast-Quality* score Q_c:

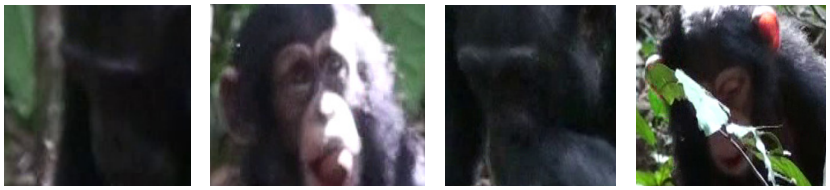
$$\sigma_{\mathbf{I}} = \sqrt{\frac{1}{h \cdot w} \sum_{x=1}^{w} \sum_{y=1}^{h} (\boldsymbol{I}(x,y) - \mu_{\mathbf{I}})^2}. \tag{4.25}$$

The maximum possible standard deviation of an image $\sigma_{\mathbf{I},\max} = \frac{1}{2} I_{\max}$ would be obtained for an equal number of black and white pixels within a binary image. Keeping in mind the naturally low contrast of chimpanzee and gorilla faces, it was empirically found that the best results can be obtained if the *Contrast-Quality* is set to 1 if $\frac{1}{4}\sigma_{\mathbf{I},\max} \leq \sigma_{\mathbf{I}} \leq \frac{1}{2}\sigma_{\mathbf{I},\max}$. Otherwise the *Contrast-Quality* is linearly decreasing and 0 if the RMS contrast is either 0 or $\frac{1}{2} I_{\max}$.

Figure 4.12 shows faces with unfavorable brightness or contrast conditions which were automatically detected by SHORE™ but might be challenging to identify. The proposed quality parameters reflect the visual quality of such images. Hence, weights can be assigned to each face within a face-track which can then be used to select the frames that are best suited for identification.

4.3.2.3. Combination of Pose and Quality Parameters

The previous sections gave an overview of the proposed parameters to estimate the quality of a facial image. However, to select the frames of a face-track which are best suited for identification a single quality measure is required which can be calculated as a combination of the parameters explained above. Figure 4.13 illustrates the process of parameter combination. Since all parameters are considered to be of equal importance, the extracted quality parameters are combined in a multiplicative fashion.



(a) (b) (c) (d)

Figure 4.12.: ***Example detections in video sequences with unfavorable illumination conditions.*** The figure shows an underexposed image (a) and an overexposed image (b). The underexposed image does not contain enough information for an accurate recognition, while the difficult lighting condition in (b) masks information which might be important for identification. Figures (c) and (d) show facial images with too low and too high contrast levels, respectively. Again, discriminative descriptors cannot be extracted from (c) due to its low contrast. On the other hand, the high contrast in (d) is due to occlusion by leafs which have a different reflectance characteristic than the actual face.

For instance, a face although in a full-frontal pose is not suited for subsequent recognition if it is blurred or underexposed. On the other hand, a facial image gathered under excellent lighting conditions cannot be used for identification if it shows a face in profile pose. Thus, since all extracted parameters are in the range $[0, 1]$, where 1 represents a high visual quality, the *Illumination-Quality* Q_i is defined by the multiplication of the three lighting parameters Q_{ue}, Q_{oe}, and Q_c

$$Q_i = Q_{ue} \cdot Q_{oe} \cdot Q_c. \tag{4.26}$$

Likewise, the *Visual-Quality* Q_v can be calculated as the product of the *Illumination-Quality* Q_i and the *Blur-Quality* Q_b as

$$Q_v = Q_i \cdot Q_b, \tag{4.27}$$

and finally the overall *Face-Quality* is given by

$$Q_f = Q_v \cdot Q_p, \tag{4.28}$$

where Q_p is the quality assigned by the Pose-Module.

The *Face-Quality* measure Q_f is finally used to select the F best frames of a face-track which are most suited for identification. Moreover, by introducing a threshold τ, face-tracks with a low maximal *Face-Quality* can be omitted from recognition which might increase the false rejection rate but on the other hand ensures a high correct classification rate. Furthermore, a low maximal *Face-Quality* measure might also help to automatically identify false positive detections which can then be omitted from further processing[7].

[7] Note that false positive detections should also be automatically rejected by the system using the confidence measures introduced in Section 4.2.3.3.

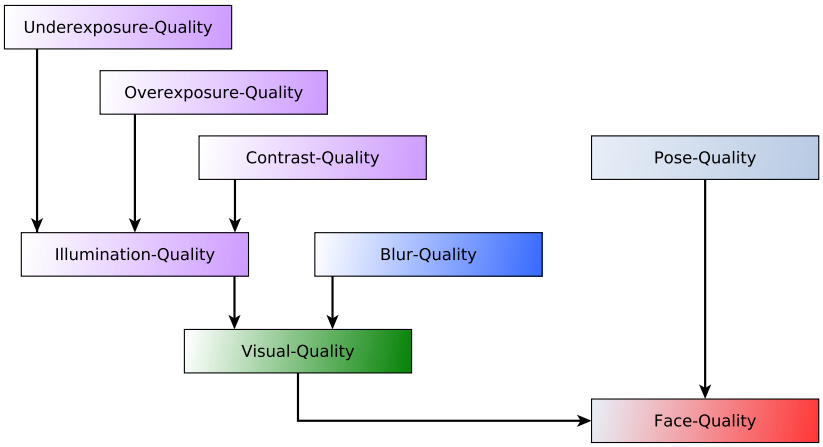


Figure 4.13.: ***Combination of estimated quality parameters to an overall face quality measure.*** The graph depicts the combination of all extracted quality parameters to a single final score which determines the overall *Face-Quality*. This measure forms the basis for selecting the best frames of a face-track which are subsequently identified by the proposed face recognition algorithm.

4.3.3. Temporal Fusion using Frame-Weighting

As outlined above, video acquisition in natural habitats of great apes often leads to large quality variations between frames. It will be shown in Section 5.5 that classification on a single frame basis usually does not lead to the desired result since it often is not distinctive enough to perform accurate identification. Thus, a novel frame-weighting approach is proposed which combines the individual frame-based classifications into a single result per face-track and hence penalizes uncertain frames.

The first step is to sort the frames of an extracted face-track according to their estimated visual quality. Based on the observation that not all frames of a given face-track are high-quality facial images and hence some frames are better suited for recognition than others, the F frames with the highest qualities are selected, aligned, and classified according to the proposed image face recognition approach described in Section 4.2.3. The following confidence measures can subsequently be derived for every classification:

- **Classification of global features**: First, three measures for the recognition pipeline using global features and SRC are proposed:
  1. *Minimal residual*:
     The minimal residual between the test image $\boldsymbol{t}$ and the matrix of training samples $\boldsymbol{A}(\boldsymbol{\delta}_i \odot \hat{\boldsymbol{p}}_1)$ (see Equation 2.29) is chosen as the first confidence measure. The smaller the minimal residual, the more confident is the classifier.

2. *Absolute difference of the first and second residual*:
   In case of misclassification, the difference of the minimal and second smallest residual is usually smaller than in the correct case. Hence, the absolute difference of the smallest two residuals Δr is used as second confidence measure. A confident classification should have a high Δr, while for incorrect classifications the difference of the two smallest residuals is rather small.

3. *Sparse Concentration Index (SCI)*:
   Besides the minimal residual, Wright *et al.* propose to utilize the sparsity of the vector $\hat{\boldsymbol{p}}_1$ as an additional confidence measure of SRC. Therefore, a measure called SCI was introduced which is defined for the sparse coefficient vector $\hat{\boldsymbol{p}}_1$ obtained from Equation 2.28 as

$$\mathrm{SCI}(\hat{\boldsymbol{p}}_1) = \frac{C \cdot \max_i \left(\|\boldsymbol{\delta}_i \odot \hat{\boldsymbol{p}}_1\|_1\right) / \|\hat{\boldsymbol{p}}_1\|_1 - 1}{C - 1} \quad \in [0, 1], \tag{4.29}$$

   where C is the number of classes and $\boldsymbol{\delta}_i$ represents the characteristic function of class i. The lager the SCI, the sparser the vector $\hat{\boldsymbol{p}}_1$ which in turn is a measure for the confidence of the classifier. For instance if $\mathrm{SCI}(\hat{\boldsymbol{p}}_1) = 1$ the test sample $\boldsymbol{t}$ is represented by only one single instance of the training set $\boldsymbol{A}$. However, if $\mathrm{SCI}(\hat{\boldsymbol{p}}_1) = 0$ the coefficient vector is extremely dense and the coefficients are equally distributed over the whole training set. Thus, the higher the SCI the more confident is the classification of the system. For the example illustrated in Figure 4.8, the SCI for the ℓ_1-norm case is $\mathrm{SCI}(\hat{\boldsymbol{p}}_1) = 0.44$ while for ℓ_2-norm minimization the resulting SCI is $\mathrm{SCI}(\hat{\boldsymbol{p}}_2) = 0.048$.

- **Classification of local features**: Secondly, two additional measures are taken into account as confidence values for the recognition pipeline using local features in combination with an SVM:

  1. *SVM probability*:
     As utilized for the proposed decision fusion scheme, the probability estimates of LIBSVM [77, 358] are used as confidence measures for identification based on local features. While for SRC the *minimal* residual determines the class affiliation, for SVM the test sample is assigned to the class with the *maximal* probability.

  2. *Difference of the highest and second highest probability*:
     Similar to the difference of the two smallest residuals, the difference between the two largest probabilities is utilized for classification via SVM. A confident classification should have a large difference between the two top scoring probabilities.

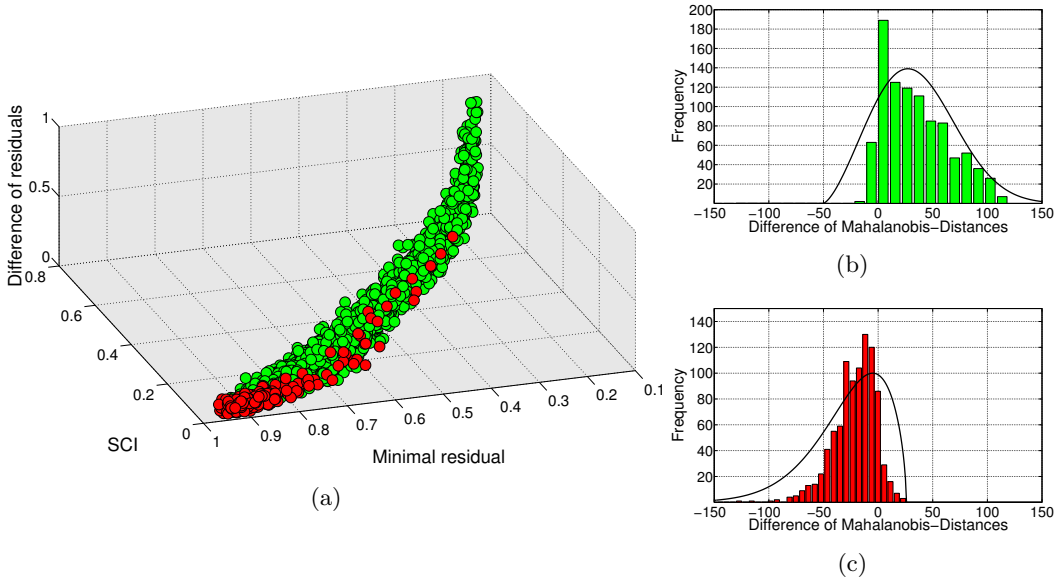


Figure 4.14.: ***Clusters of correct and incorrect classifications as well as the histograms of Mahalanobis-distances.*** Figure (a) depicts the proposed confidence vectors of correct classifications (green samples) and incorrect classifications (red samples). Note that for illustration purposes only the first three dimensions of the original 5-D space is plotted. It can be seen that the incorrect classifications are mostly located in one corner while the cluster of correct classifications is more scattered in space. Figures (b) and (c) illustrate the histograms of the difference of Mahalanobis-distances $\Delta d_M(\boldsymbol{t})$ for correct and incorrect classifications, respectively. The fitted Extreme Value Distributions(EVDs) are superimposed in black.

All five measures are subsequently concatenated into a confidence vector $\boldsymbol{v}$. The goal of the proposed frame-weighting approach is to estimate the probability that the classification of frame f was correct. This probability serves as a weighting factor which is subsequently assigned to each classified frame in order to obtain a final prediction. The proposed frame weighting approach is thus divided into a *training* and a *test* phase:

The Training Phase: First, a 20-fold Monte-Carlo cross-validation is applied on the training set in order to construct the confidence-vectors for each correct and incorrect classification. This procedure results in two clusters of correct and incorrect classifications in 5-dimensional space. Figure 4.14(a) depicts a scatter plot of the two resulting clusters, where the red samples denote misclassifications and the green circles represent correctly classified samples.

Misclassifications are mostly located within a narrow area while the cluster of correct classifications is more scattered in space. However, it can be seen that correctly classified samples can be separated from misclassifications quite well since only a few incorrectly identified samples are located in the area of the majority of correct classifications.

In order to estimate the probability of a correct classification, the next step is to calculate the Mahalanobis-distances [376] of each sample to the cluster of correct classifications and to the cluster of false classifications. The Mahalanobis-distance of a test vector $\boldsymbol{t}$ to a group of observations $\boldsymbol{O}$ with mean-vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{S}$ is defined as

$$d_M(\boldsymbol{t}, \boldsymbol{O}) = \sqrt{(\boldsymbol{t} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{S}^{-1} (\boldsymbol{t} - \boldsymbol{\mu})}. \tag{4.30}$$

The difference of Mahalanobis-distances $\Delta d_M(\boldsymbol{t})$ is thus given by

$$\Delta d_M(\boldsymbol{t}) = d_M^{IC}(\boldsymbol{t}, \boldsymbol{O}_{IC}) - d_M^{C}(\boldsymbol{t}, \boldsymbol{O}_{C}), \tag{4.31}$$

where $d_M^{IC}(\boldsymbol{t}, \boldsymbol{O}_{IC})$ is the Mahalanobis-distance of $\boldsymbol{t}$ to the cluster of incorrect classifications $\boldsymbol{O}_{IC}$ and $d_M^{C}(\boldsymbol{t}, \boldsymbol{O}_{C})$ represents the Mahalanobis-distance of the test sample to the cluster of correct classifications. This procedure is repeated in a leave-one-out fashion, i.e. every correctly and incorrectly classified sample takes the role of the test vector $\boldsymbol{t}$ once. Figures 4.14(b) and 4.14(c) exemplarily show the obtained histograms of $\Delta d_M(\boldsymbol{t})$ for correct and incorrect classifications, respectively.

The next step of the training phase is to apply Maximum Likelihood Estimation (MLE) in order to fit an Extreme Value Distribution (EVD) [377] to the obtained histograms. Note that only the key-steps of estimating the parameters of an EVD are illustrated below. For a more detailed derivation of the subsequent equations the interested reader is referred to [378].

The Probability Density Function (PDF) of the EVD given an extreme-valued random variable x is defined as

$$P(x) = \lambda \exp\left[-\lambda(x - \mu) - \mathrm{e}^{-\lambda(x-\mu)}\right], \tag{4.32}$$

where μ and λ are location and scale parameters, respectively. The likelihood of drawing N samples x_i from an EVD with parameters μ and λ can thus be written as

$$P(x_1 \ldots x_N | \lambda, \mu) = \prod_{i=1}^{N} \lambda \exp\left[-\lambda(x_i - \mu) - \mathrm{e}^{-\lambda(x_i-\mu)}\right]. \tag{4.33}$$

Maximizing Equation 4.33 with respect to λ and μ yields the maximum likelihood estimation of both parameters.

From a mathematical perspective it is often more convenient to work with the logarithmized version of Equation 4.33 known as the *log-likelihood-function*:

$$\mathcal{L}(\mu, \lambda) = \log(P(x_1 \ldots x_N | \lambda, \mu)) = N \log(\lambda) - \sum_{i=1}^{N} \lambda(x_i - \mu) - \sum_{i=1}^{N} \mathrm{e}^{-\lambda(x_i-\mu)} \tag{4.34}$$

In order to find the parameters $\hat{\lambda}$ and $\hat{\mu}$ that maximize the log-likelihood function, the next step is to get the partial first derivatives of $\mathcal{L}(\mu, \lambda)$ and set them to zero:

$$\frac{\partial \mathcal{L}(\mu, \lambda)}{\partial \mu} = N\lambda - \lambda \sum_{i=1}^{N} \mathrm{e}^{-\lambda(x_i - \mu)} \stackrel{!}{=} 0 \tag{4.35}$$

$$\frac{\partial \mathcal{L}(\mu, \lambda)}{\partial \lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N}(x_i - \mu) + \sum_{i=1}^{N}(x_i - \mu)\mathrm{e}^{-\lambda(x_i - \mu)} \stackrel{!}{=} 0. \tag{4.36}$$

After a few algebraic steps it can be shown from Equation 4.35 that $\hat{\mu}$ is given by

$$\hat{\mu} = -\frac{1}{\hat{\lambda}} \log \left[\frac{1}{N} \sum_{i=1}^{N} \mathrm{e}^{-\hat{\lambda} x_i} \right]. \tag{4.37}$$

Substituting Equation 4.37 into Equation 4.36 finally yields

$$\mathcal{K}(\hat{\lambda}) = \frac{1}{\hat{\lambda}} - \frac{1}{N} \sum_{i=1}^{N} x_i + \frac{\sum_{i=1}^{N} x_i \mathrm{e}^{-\hat{\lambda} x_i}}{\sum_{i=1}^{N} \mathrm{e}^{-\hat{\lambda} x_i}} \stackrel{!}{=} 0. \tag{4.38}$$

In order to find the root of Equation 4.38, the *Newton-Raphson-Algorithm* [379, 380] can be applied for which the first derivative is needed:

$$\frac{d\mathcal{K}}{d\hat{\lambda}} = -\frac{1}{\hat{\lambda}^2} + \left(\frac{\sum_{i=1}^{N} x_i \mathrm{e}^{-\hat{\lambda} x_i}}{\sum_{i=1}^{N} \mathrm{e}^{-\hat{\lambda} x_i}} \right)^2 - \frac{\sum_{i=1}^{N} x_i^2 \mathrm{e}^{-\hat{\lambda} x_i}}{\sum_{i=1}^{N} \mathrm{e}^{-\hat{\lambda} x_i}}. \tag{4.39}$$

The Newton-Raphson-Algorithm which finds the root of Equation 4.38 and consequently estimates $\hat{\lambda}$ is summarized in Algorithm 2.

Once $\hat{\lambda}$ has been estimated it can be plugged into Equation 4.37 in order to calculate $\hat{\mu}$. The fitted EVDs of the difference of Mahalanobis-distances of correct and incorrect classifications are superimposed in Figures 4.14(b) and 4.14(c), respectively. As can be seen, both histograms can be well approximated by an EVD.

Algorithm 2 Newton-Raphson-Algorithm which finds the root of Equation 4.38

Require: $\hat{\lambda}_0$: Initial random guess of $\hat{\lambda}$
ϵ: Threshold to stop while-loop
M : Maximum number of iterations

function NEWTON-RAPHSON-ALGORITHM($\hat{\lambda}_0, \epsilon, M$)

$It \leftarrow 0$ ▷ Set number of iterations to zero
while ($\mathcal{K}(\hat{\lambda}) \geq \epsilon$) **and** ($It < M$) **do**
$\hat{\lambda} \leftarrow \hat{\lambda}_0 - \frac{\mathcal{K}(\hat{\lambda}_0)}{\mathcal{K}'(\hat{\lambda}_0)}$ ▷ Compute new estimate of $\hat{\lambda}$
$\mathcal{K}(\hat{\lambda}) \leftarrow \frac{1}{\hat{\lambda}} - \frac{1}{N}\sum_{i=1}^{N} x_i + \frac{\sum_{i=1}^{N} x_i \mathrm{e}^{-\hat{\lambda}x_i}}{\sum_{i=1}^{N} \mathrm{e}^{-\hat{\lambda}x_i}}$ ▷ Plug current $\hat{\lambda}$ into Equation 4.38
$It \leftarrow It + 1$ ▷ Increment iteration
$\hat{\lambda}_0 \leftarrow \hat{\lambda}$ ▷ Update $\hat{\lambda}_0$
end while
print 'Approximation of $\hat{\lambda}$ is: ',$\hat{\lambda}_0$
if $\mathcal{K}(\hat{\lambda}_0) \geq \epsilon$ **then**
print 'Required accuracy not reached in ',M,' iterations!' ▷ Warning
end if
end function

The Test Phase: After parameter estimation using MLE, the fitted EVDs can be utilized in the test phase to calculate the probability that the classification in frame f was correct and weight the resulting prediction accordingly.

The weighting-factor w_f of frame f is given by the Bayes' theorem. Let $P(\Delta d_M|C)$ be the probability of the difference of Mahalanobis-distances given a correct classification, then the weighting-factor $w_f = P(C|\Delta d_M)$ is given by

$$P(C|\Delta d_M) = \frac{P(\Delta d_M|C)P(C)}{P(\Delta d_M)}, \tag{4.40}$$

where $P(C)$ is the probability of a correct classification and $P(\Delta d_M)$ is the probability of the difference of Mahalanobis-distances. The correct classification probability $P(C)$ is taken from the accuracy of the proposed system after 20-fold crossvalidation applied on the training set as explained above. Moreover, $P(\Delta d_M|C)$ is calculated from the estimated PDF of correct classifications. By applying the *law of total probability*, $P(\Delta d_M)$ can be expanded which yields the following reformulation of the Bayes' law:

$$P(C|\Delta d_M) = \frac{P(\Delta d_M|C)P(C)}{P(\Delta d_M|C)P(C) + P(\Delta d_M|IC)P(IC)} \\ = \frac{P(\Delta d_M|C)P(C)}{P(\Delta d_M|C)P(C) + P(\Delta d_M|IC)(1 - P(C))}, \tag{4.41}$$

where $P(IC) = 1 - P(C)$ is the probability of an incorrect classification and $P(\Delta d_M|IC)$ is the probability of the difference of Mahalanobis-distances given an incorrect classification. The conditional probability $P(\Delta d_M|IC)$ is estimated from the PDF of the Extreme Value Distribution (EVD) of incorrect classifications fitted during training.

Once the weighting factor $w_f = P(C|\Delta d_M)$ has been calculated for every selected frame, the frame-weighting procedure is as follows: Let $\boldsymbol{s}_{\mathrm{df},f} \in \mathbb{R}^{C\times 1}$ be the score vector introduced in Section 4.2.3.3 for frame f, then the cumulative score vector $\boldsymbol{s}_c \in \mathbb{R}^{C\times 1}$ is defined as the weighted average of the score-vectors for all selected frames $f = 1 \dots F$

$$\boldsymbol{s}_c = \frac{1}{F} \sum_{f=1}^{F} w_f \cdot \boldsymbol{s}_{\mathrm{df},f}. \tag{4.42}$$

Once all frames have been processed, the index of the maximum element of $\boldsymbol{s}_c$ denotes the final prediction of the current face-track $\boldsymbol{t}_{\mathrm{Track}}$.

$$\mathrm{ID}(\boldsymbol{t}_{\mathrm{Track}}) = \arg\max_i \{\boldsymbol{s}_c(i)\}. \tag{4.43}$$

4.4. Chapter Summary

During the course of this chapter a detailed explanation of the proposed Primate Recognition Framework (PRF) for detection and individual identification of great apes in natural habitats was given. Since humans and great apes share similar properties of their facial appearance, identification algorithms originating from human face recognition techniques were adapted and extended to robustly recognize primates in real-world environments. The developed framework thus comprises three main parts: *detection*, *alignment*, and *identification*.

The first part of this chapter introduced algorithms for detection and recognition of great apes in still images. The third-party library SHORE™, developed by Fraunhofer IIS, is utilized to robustly detect faces and facial keypoints of chimpanzees and gorillas. Based on the detected landmarks, an *affine transformation* is applied to ensure that facial features such as eyes, nose, and mouth are located at the same positions throughout the entire dataset. This procedure ensures comparability of extracted visual descriptors for all faces. Subsequently, a global descriptor based on a feature-level combination of Gabor-wavelets and Extended Local Ternary Patterns (ELTPs) as well as local SURF features extracted around six facial interest points are calculated after alignment and scaling. Thus, not only the holistic appearance of a primate face is utilized for identification, but also discriminating wrinkle patterns around both eyes and the nose of great apes are used implicitly. Morevover, a simple but effective decision-level fusion method was proposed to combine the results of global and local features.

The second part of this chapter extended these ideas to identify primates in videos. A requirement of robust identification of great apes in video sequences is to be able to track detected faces through the scene. Again SHORE™ is utilized for that purpose. Since the visual quality of the first frame of an extracted face-track is often not suited well enough for robust and accurate identification, several quality assessment modules were proposed, including pose, blur, and illumination estimation. Thereby, the frames of face-tracks can be sorted according to their visual quality. The F frames which are best suited for subsequent identification are selected which significantly increases the accuracy of the system as will be shown in Section 5.5. Finally, a novel frame-weighting approach was proposed which probabilistically weights the predictions obtained from different frames by taking various confidence measures into account.

5. Evaluation and Results

5.1. Chapter Overview

In this chapter the proposed Primate Recognition Framework (PRF) is thoroughly evaluated on realistic datasets of free-living and captured primate individuals. Many databases which are commonly used as benchmark datasets for human face recognition were compiled under laboratory conditions. Contrarily, the datasets used in this thesis were gathered in real-world environments and are extremely challenging since they comprise a variety of uncontrollable extrinsic and intrinsic factors which are frequently present in visual footage gathered under natural conditions. Thus, the datasets used for experimentation in this thesis place high demands on facial recognition systems since they realistically reflect challenges present in real-world environments. In order to evaluate the performance of the proposed framework, it is compared against other state-of-the-art face recognition approaches that were originally developed to identify humans based on their facial appearance.

Section 5.2 introduces the annotation tool which was used by experts to annotate image data for training and experimentation. Subsequently, the datasets of captive and free-living chimpanzee and gorilla individuals are described in detail. **Section 5.3** reviews the evaluation measures which are commonly used in the research community to compare face recognition algorithms and asses their performances. **Section 5.4** presents the results of the proposed PRF for the identification of primates in still images. The developed non-invasive identification system is benchmarked against eight different state-of-the-art approaches for human face recognition. Furthermore, different face alignment and image enhancement techniques are applied to measure the influence of miscellaneous pre-processing methods. The proposed global and local face recognition pipelines are first evaluated separately in a closed-set identification scheme. Secondly, it is shown that the proposed decision fusion scheme can further enhance the performance of the system. After thorough evaluation on still images, it is investigated in **Section 5.5** how the proposed framework performs when applied to video sequences. Here, additional recording artifacts and coding errors such as interlacing, blur, and blocking might hamper classification. It is shown how the proposed pose estimation and quality assessment modules can be applied in order to find the frames within a face-track which are best suited for identification. Moreover, it is demonstrated that weighting the contributions of multiple frames using the proposed frame-weighting scheme improves the overall performance in comparison to traditional weighting paradigms and approaches that rely solely on the classification of a single frame.

5.2. Annotation and Description of Datasets

5.2.1. The Annotation Tool

To provide a valid ground-truth database of facial images which can be used for training and evaluation, a proprietary annotation tool developed by Fraunhofer IIS called *ImageMarker* [381] was extended and adapted for this purpose. A screenshot of a sample annotation using the *ImageMarker* can be seen in Figure 5.1.

Figure 5.1.: ***Screenshot of the* ImageMarker *for ground-truth annotation.*** (1) All faces in the image are marked with blue rectangles and the position of facial marker points such as eyes, mouth, as well as earlobes have to be annotated. (2) Additional metadata such as the species, gender, identity, and others are assigned to each face. (3) Furthermore, the head pose of the primate face can be annotated. (4) Once all faces are marked and equipped with attributes, the annotations are automatically saved into an XML-file and the next image in the list can be annotated. Note that if the operator is unsure about a certain attribute it can also be annotated as *unknown*. [I04]

Experts in the field of animal monitoring and biodiversity conservation annotated the location of faces as well as facial marker points which are important to subsequently align the images (see 1 in Figure 5.1). Moreover, additional metadata in form of key-value pairs such as pose, gender, species, age as well as the name of the individual were assigned to every annotated face (see 2 and 3 in Figure 5.1). The output of the annotation process is an XML-file for every image which contains information about all objects within the photograph. Thereafter, for training and benchmarking of the face detection and identification algorithms, facial images can be selected and cropped according to the given annotations. Figure 5.2 shows the XML-scheme of the provided ground-truth data generated by the *ImageMarker* by Fraunhofer IIS.

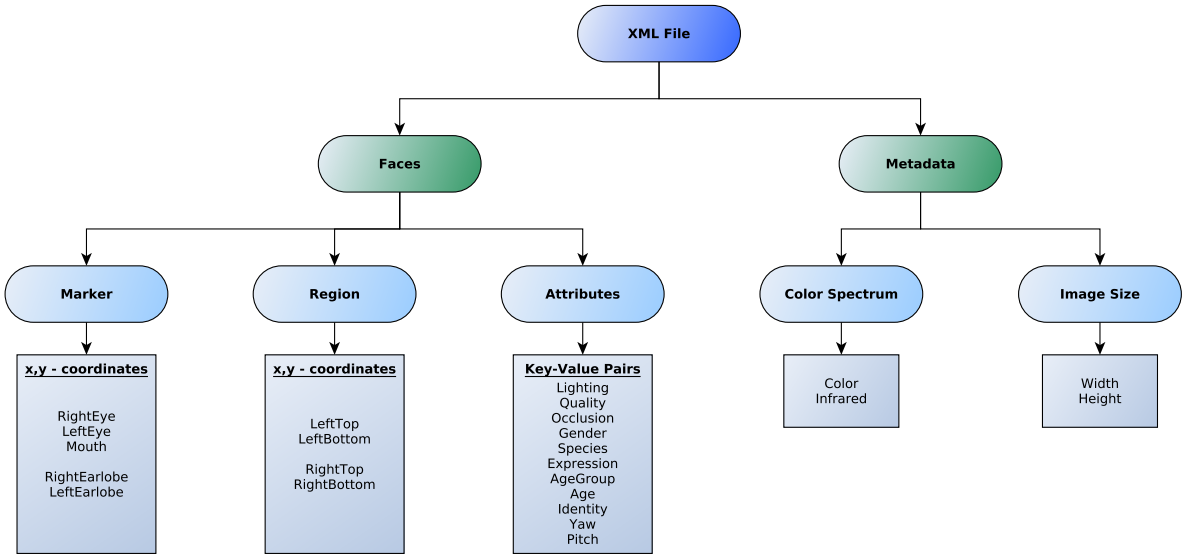


Figure 5.2.: ***XML-scheme of the ground-truth data generated by the*** **ImageMarker.** In addition to general information about the original image such as the *Color Spectrum* and the *Image Size*, every annotated face within an image is assigned with information about the *Region* of interest, certain *Marker* points, as well as a number of facial *Attributes* in form of key-value pairs.

Besides information about the original image itself such as the *Color Spectrum* (color or infrared image) and its size, any annotated face within the photograph is labeled with three fields: *Marker*, *Region*, and *Attributes*. The *Marker* field contains the x, y-coordinates of the facial features. For full-frontal faces, the coordinates of both eyes and the mouth were annotated. For faces in profile views either the left or the right earlobe was annotated depending on the pose of the head. The *Region* field on the other hand contains the coordinates of the upper left as well as the lower right corner of the rectangle drawn around the primate faces which can later be used to cut the regions of interest out of the images. Furthermore, key-value pairs were assigned to every annotated face and saved within the *Attributes* field which allows automatic selection of images according to various parameters.

All possible key-value pairs of the annotated attributes are described in Table 5.1.

Attributes	Possible Values	Description
Lighting	Dark, HalfLight, Light	Annotation of the current lighting situation with respect to overexposure and underexposure.
Quality	Poor, Fair, Good	The overall quality of the facial image which is influenced by blur or noise for instance.
Occlusion	None, Moderate, Much	Annotation of the level of occlusion: Small occlusions by branches or leafs for instance are annotated as *Moderate* while severe occlusions that overlay more than half of the face are annotated as *Much*.
Gender	Female, Male	The gender of the individual.
Species	Chimp, Gorilla	Annotation of the species: The user currently has to decide only between chimpanzee and gorilla.
Expression	RelaxedLip, Playing, Pout, BaredTeeth, Eating	Categorization of facial expressions of great apes into five different taxonomies which are commonly used by biologists and behavioral scientists.
AgeGroup	Infant, Juvenile, SubAdult, Adult, Elderly	Rough estimation of the age group: For individuals living in their natural habitats the exact age is often not known.
Age	—	The age of an individual: This is only annotated if the true age is known to the annotator.
Identity	—	The name of the individual.
Yaw	Left, SemiLeft, Front, SemiRight, Right	The horizontal face pose also referred to as yaw.
Pitch	SemiTop, EyeLevel, SemiBottom	The vertical pose of the primate face also known as pitch.

Table 5.1.: ***Description of the key-value pairs used for annotations.*** The table gives an overview of all possible key-value pairs used for the annotations of the datasets. These attributes can be used to filter the datasets and test the robustness of developed algorithms against different poses, lighting conditions, various degrees of occlusion, and other challenges present in real-world environments.

In case an annotator is uncertain about an attribute it can also be annotated as *"unknown"*. Note that for benchmark purposes, the datasets used for evaluation in this thesis can be requested via `http://www.saisbeco.com`.

5.2.2. Image Datasets of Captive and Free-living Primate Individuals

Due to the lack of publicly available face databases of great apes, self-assembled annotated datasets of captive as well as free-living individuals from the zoo of Leipzig, Germany and the Taï National Park, Côte d'Ivoire, Africa are used for benchmark purposes within this thesis. This section gives an overview of the datasets of captive and as free-living chimpanzees and gorillas which were used for evaluation of the proposed PRF.

Captive chimpanzees and gorillas: The study subjects for the first dataset of captive individuals are 24 chimpanzees (*Pan troglodytes*) separated into two groups from the zoo of Leipzig, Germany. Video material from each individual was collected between June 2010 and December 2010. A high-definition camcorder (Sony Handycam, 3.1 Mpx, 25×optical zoom) with a tripod was placed on one of five observation platforms from which a barrier-free view down into the enclosures of the study groups was guaranteed. Individuals were recorded for one to five minutes in order to capture different expressions, poses, and varying lighting conditions. Finally, snapshots of the gathered video material were extracted and annotated to provide a realistic and valid dataset for training and evaluation. Furthermore, a second dataset of captive western gorillas (*Gorilla gorilla*) was collected with the same acquisition protocol as explained above. These datasets are referred to as *ChimpZoo* and *GorillaZoo* for the remainder of the thesis. To give an impression of both datasets, example images of captive chimpanzees (a) and gorillas (b) are illustrated in Figure 5.3.

(a) ChimpZoo

(b) GorillaZoo

Figure 5.3.: ***Example images of captured chimpanzees and gorillas.*** (a) Images of the ChimpZoo dataset. (b) Photographs of the GorillaZoo dataset. Both datasets were gathered at the zoo of Leipzig, Germany. [I04]

Free-living chimpanzees: In order to proof the applicability of the developed PRF to data gathered in natural habitats, additionally a third dataset of free-living chimpanzees was assembled. Video footage of a group of chimpanzees (*Pan troglodytes*) was gathered at the Taï National Park, Côte d'Ivoire, Africa. Pictures were taken using a Canon EOS D30 (3 Mpx) camera between November 2003 and December 2004. The dataset is referred to as *ChimpTaï* for the remainder of this thesis. Figure 5.4 shows example snapshots of the *ChimpTaï* dataset.

Figure 5.4.: ***Example images of free-living chimpanzees and gorillas.*** Images of the ChimpTaï dataset gathered at the Taï National Park, Côte d'Ivoire, Africa. [I02]

It is obvious that all three datasets are extremely challenging because detected faces of one single individual might appear in a variety of different poses, expressions, variation of illumination and even partial occlusion by branches, leafs, or conspecifics. Thus, the algorithms used for identification need to be robust against that kind of variations in order to achieve sufficient recognition accuracy. Details about all datasets are listed in Table 5.2.

Dataset	Images	Individuals	Faces	Frontal Faces	Faces per Individual	Frontal Faces per Individual
ChimpZoo	2118	24	2438	648	75 – 128	14 – 43
GorillaZoo	499	6	540	154	64 – 116	19 – 33
ChimpTaï	3905	80	5079	1294	1 – 342	1 – 84

Table 5.2.: ***Details about the datasets* ChimpZoo, GorillaZoo, *and* ChimpTaï.** Note that the columns *Frontal Faces* and *Frontal Faces per Individual* refer to facial images that were annotated as *Front* with arbitrary vertical directions (pitches).

5.2.3. Generation of Datasets for Experiments and Evaluation

Datasets gathered in real-world environments often have two major drawbacks. First, as can be seen in Table 5.2, many classes do not contain enough images to build a sufficiently representative feature space in order to learn its intrinsic structure. Secondly, facial images might carry a large variety of different poses, lighting conditions, occlusions, expressions and many more challenging extrinsic and intrinsic factors.

It is generally agreed that a sufficiently constrained pattern recognition problem, i.e. low intra-class variance and high inter-class variance, can be solved more efficiently and with a much higher confidence than ill-defined problems [382]. On the other hand, like every machine learning algorithm the fundamental goal of the proposed PRF is to generalize beyond the examples of the training set. Thus, training an algorithm solely on unrealistic data will lead to a system that is prone to challenges present in real-world application scenarios. Therefore, dataset generation used for experimentation in this thesis are based on the following criteria. This procedure ensures that individual models are trained on data that is suited for the task at hand, so that the intrinsic structure of the feature space can be learned sufficiently well. At the same time it is ensured that the system is evaluated on realistic data which can be expected in real-world applications:

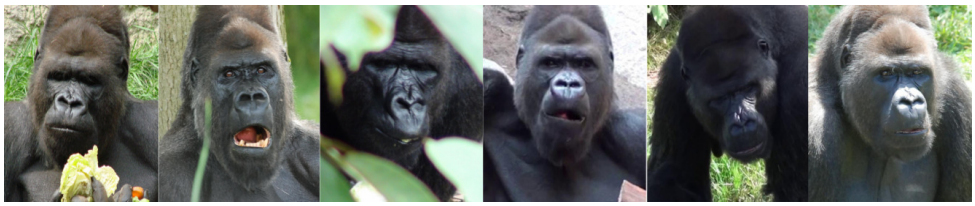
- **Number of images per individual:** In order to generate a feature space suitable for recognition and subspace learning, all individuals with less than 5 images in the database are excluded from further processing.
- **Size:** Only face images that provide sufficient facial details in order to extract discriminating descriptors should be part of the evaluation data. Hence, only faces with a size of at least 64×64 pixels are considered to be in the dataset.
- **Vertical pose (pitch):** Although too large pitch angles might influence the recognition results negatively, primate faces in natural environments often appear in a large variety of vertical poses. Therefore, images with all possible annotated pitch angles, i.e. *EyeLevel*, *SemiBottom*, and *SemiTop* are considered.
- **Horizontal pose (yaw):** Since the proposed PRF is designed for full-frontal head poses, faces with profile views are excluded from the training set and thus only faces, annotated as *Frontal* are selected. However, as shown in Section 5.4.3, the proposed face recognition algorithm is also tested on semi-profile face images in order to evaluate the algorithms performance on non-frontal data.
- **Occlusion:** Also too much occlusion by branches, leafs or other individuals might hamper training and classification. Therefore, only face images with *None* or *Moderate* occlusion are taken into account since faces with too much occlusion will not even be detected by SHORE™ and thus do not need to be identified.
- **Quality:** Although the overall image quality can significantly hamper the performance of the system, all annotated quality parameters are considered for the generation of the datasets to ensure a realistic evaluation scenario.

- **Facial Expression:** Since faces are deformable objects, they can significantly change their appearance with different facial expressions. Thus, face images with a high degree of facial expressions might be extremely challenging to classify. In real-world scenarios, however, facial expressions occur frequently while eating, playing, or shouting. Hence, all annotated facial expressions are considered to be in the evaluation sets.
- **Unknown Identities:** Moreover, facial images for which the annotator was not sure about the name of the individual are excluded from the datasets to ensure validity of the experiments conducted in this thesis.

Figure 5.5 shows detected faces of one individual for the ChimpZoo (a), ChimpTaï (b), and GorillaZoo dataset (c), respectively.

(a) ChimpZoo

(b) ChimpTaï



(c) GorillaZoo

Figure 5.5.: ***Facial images of one individual per dataset.*** The figure shows different facial appearances of one individual for the (a) ChimpZoo dataset, (b) ChimpTaï dataset, and (c) GorillaZoo dataset, respectively. Note that the face of a single individual can show a variety of different appearances due to various lighting conditions, facial expressions, head poses, and even partial occlusion by branches, leafs, or conspecifics. All these factors place high demands on the identification system.

It is obvious that all datasets are extremely challenging for the recognition task because detected faces of one single individual can show a variety of different appearances due to expressions, lighting conditions, and even partial occlusion by branches, leafs, or other objects. Furthermore, even for primate faces annotated as *Frontal* a significant amount of variation is present unlike for most human face datasets such as the *Extended YaleB* [383] or the *Color FERET* [384] database. Thus, the algorithms used for identification are required to be robust against that kind of variations in order to achieve sufficient recognition results. Table 5.3 gives a detailed overview of all four filtered datasets used for evaluation.

Dataset	Number of Images	Number of Individuals	Frontal Faces per Individual
ChimpZoo	572	24	$14 - 38$
GorillaZoo	133	6	$16 - 28$
ChimpTaï	988	49	$5 - 60$

Table 5.3.: ***Details about the filtered datasets used for the experiments within this thesis.*** The table lists the details of the filtered versions of the datasets *ChimpZoo*, *GorillaZoo*, and *ChimpTaï* used for the experiments conducted in this thesis.

Although the developed primate face recognition system is mainly designed to recognize full-frontal faces, it is crucial to evaluate the system's robustness against different poses. Using the annotations outlined in the previous chapter, three distinct pose-specific subsets can be generated for every primate face database. The subset *Front* then only contains full-frontal face images of every individual, while the subsets *SemiLeft* and *SemiRight* contain images of semi-left and semi-right faces, respectively. Example images of one individual per dataset with *SemiLeft*, *Frontal*, and *SemiRight* poses can be seen in Figure 5.6.

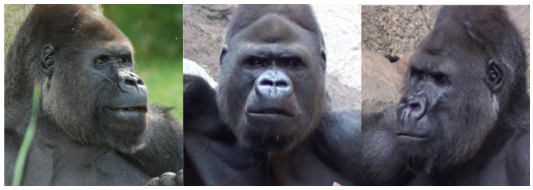
Conclusively, it can be stated that the developed PRF is thoroughly evaluated on realistic data which can be expected for real-world applications in the field. Therefore it can be assumed that the results achieved on that data can also be expected in real-world scenarios. Table 5.4 shows an overview of the pose specific subsets for all databases of captive and free-living primate individuals.

Dataset	Individuals	Frontal Faces	SemiLeft Faces	SemiRight Faces
ChimpZoo	24	572	490	506
GorillaZoo	6	133	118	98
ChimpTaï	49	988	808	831

Table 5.4.: ***Details about the off-frontal subsets.*** The table lists an overview of the off-frontal pose subsets for the the datasets *ChimpZoo*, *GorillaZoo*, and *ChimpTaï*.

(a) ChimpZoo

(b) ChimpTaï



(c) GorillaZoo

Figure 5.6.: ***Facial images taken from different pose subsets for one individual per dataset.*** The figure shows images taken from the three different pose subsets *SemiLeft* (leftmost), *Front* (middle), and *SemiRight* (rightmost) for one individual per dataset. Images were taken from the (a) ChimpZoo, (b) ChimpTaï, and (c) GorillaZoo dataset.

5.2.4. Synthesis of Unseen Image Conditions

As stated in Section 5.2.3, the minimum number of images per individual is required to be at least five in order to generate a feature space that is suitable for recognition and subspace learning. However, the elusive and uncooperative nature of great apes in real-world settings limits available data. Consequently, for the datasets gathered in natural environments, a number of individuals are omitted from analysis due to insufficient data (see Tables 5.2 and 5.3). To overcome this limitation, manually collecting and analyzing field data for individuals might be minimized by synthesizing unseen image conditions. This not only allows a more complete coverage of image parameter variability for training but could also lead to inclusion of more individuals as less real images are required.

Within a collaboration with the *Visual Information Laboratory, University of Bristol, UK*[1] preliminary experiments were conducted in [10] to explore face recognition accuracy on images of chimpanzees by synthesizing data from a generic 3D model. Automatically annotated synthetic images generated under controlled conditions by the University of Bristol were used to evaluate the performance of the proposed PRF under different poses and lighting situations.

[1] `http://www.bristol.ac.uk/vi-lab/` Last visit: August 12^{th}, 2014

Furthermore, synthetic data was used to supplement real data in order to improve the generalization capabilities of the algorithm and therefore increase the system's robustness against non-frontal posed test samples.

Although 2D images rendered from 3D models have been successfully used to improve human face recognition systems, generation of 3D models usually requires cooperation of the individuals to be scanned [385, 386]. Recently, methods were developed which use 3D modeling and synthesis without the need for scanned images. However, these approaches require several images in controlled poses [387], frontal and profile mugshots [388] or additional data such as depth maps [389].

Uncooperative subjects and images gathered under uncontrolled conditions make those approaches infeasible for great apes. Therefore, a generic shape model was used to approximate the 3D geometry of a chimp's face. The pose and lighting of a face can be varied independently. Admittedly, such a model is constrained due to limited subject information but assumed to be sufficiently representative to generate artificial test and training data in order to evaluate face recognition algorithms. The 572 frontal face images of the *ChimpZoo* dataset described in Section 5.2.3 serve as basis from which images with different vertical and horizontal poses as well as different lighting conditions were synthesized. In order to project the source images onto the surface of the model, face images first have to be optimally aligned to the 3D model. The minimal Procrustes distance [390, 391] based on the three manually annotated facial marker points described in Section 5.2.1 was applied to obtain a scale, shift, and rotation normalized texture map. Lambertian reflection [392] was utilized as a model for diffuse lighting since more complex models developed for humans [393] are not directly transferable to great apes.

By altering either pose or lighting, synthetic images were created from each of the 572 face images of the *ChimpZoo* dataset. Horizontal and vertical pose angles are each varied between $\pm 30^\circ$ with a step size of 10°. Lighting conditions are varied by changing either the position or the intensity of a virtual spotlight with one of three possible intensities: low, mid, or high. The spotlight position is varied in 30° increments between $\pm 60^\circ$. Additionally an ambient light setting is included. Since the purpose is to investigate the influence of pose and lighting changes separately, only one parameter is varied at a time and the other one remains in a "neutral" state of frontal pose or ambient lighting. Rendering under these conditions results in 5 spotlight positions, 3 exposure intensities, 7 horizontal and vertical poses for each of the 572 basis images. Therefore, in summary the final rendered dataset contains $11,440$ unique synthetic images, each containing either horizontal or vertical pose variation, spotlight exposure, or position variation. Figure 5.7 displays the variation of synthesized images generated from a single chimp face input image.

Figure 5.7.: ***Example images of synthesized data.*** The figure shows synthesized example images for a basis image taken from the *ChimpZoo* dataset. Each artificial image contains either horizontal or vertical pose variation, respectively spotlight exposure or variation of the spotlight position. [I05]

5.2.5. Video Datasets of Captive and Free-living Chimpanzees

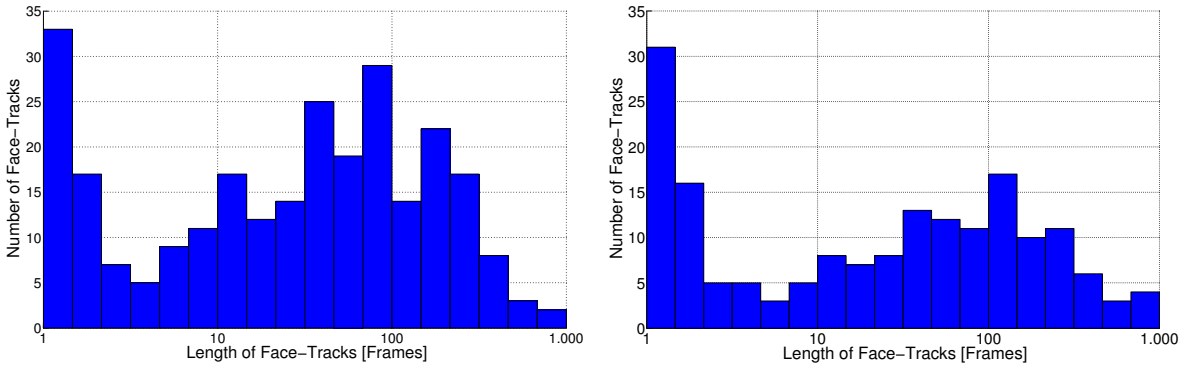
Besides evaluation on image databases, it is also investigated in this thesis how the proposed PRF performs when applied to video sequences. Thus, video footage of captured as well as free-living chimpanzees was recorded at the zoo of Leipzig, Germany and the Taï National Park, Côte d'Ivoire, Africa, using the same acquisition protocol as explained above. Since image and video datasets were gathered independently, the image databases described above can be utilized for training while the gathered video material serves as test data. Note however that not every individual of the according image dataset is necessarily also present in one of the video sequences. However, the system is trained with the entire set of individuals for each database. Hereafter, both video datasets are referred to as *ChimpZoo-Video* and *ChimpTaï-Video*, respectively.

First, a deinterlacing filter[2] was applied in order to enhance the quality of the video recordings. SHORE™ was used for annotation of the gathered video footage in order to detect and track faces of chimpanzees in a first step. Every extracted face-track is subsequently annotated with the name of the according individual. False positive detections were labeled as *unknown*. These face tracks serve as test instances for rejection, i.e. it is evaluated how well the identification module is capable to classify false positive detections as unknown and therefore reject them as non-faces. Statistics about both video datasets can be found in Table 5.5 and Figure 5.8.

[2]The *yadif*-filter of FFmpeg version 2.2.4 was used. `https://www.ffmpeg.org/` Last visit: August 12^{th}, 2014

Dataset	Videos	Face-Tracks	Frames per Face-Track	Individuals	Unknowns
ChimpZoo-Video	14	264	1 – 818	16	61
ChimpTaï-Video	11	198	1 – 1149	17	49

Table 5.5.: ***Details about the video datasets used for evaluation within this thesis.*** Face-tracks are first extracted for every video sequence using SHORE™ and subsequently labeled with the name of the according individual. False positive detections are labeled as *unknown*.



(a) ChimpZoo-Video (b) ChimpTaï-Video

Figure 5.8.: ***Histogram of the length of the extracted face-tracks.*** In both datasets most face-tracks contain a low number of frames which suggest that many faces were initially detected by SHORE™ but could not be tracked through the video sequence.

As can be seen in Figure 5.8, a relatively high number of face-tracks contain only few frames which suggests that often faces were detected but could not be tracked through the video. Many of those short face-tracks contain false positive detections. Therefore, face-tracks that are too short are ignored from further processing and are directly classified as unknown since it is assumed that these detections do not show an actual face. Details about that procedure are given in Section 5.5.

5.3. Evaluation Measures and Experimental Design

The performance of the proposed PRF is reported on two standard tasks commonly used for face recognition and biometric identification: *closed-set identification* and *open-set identification*, where the latter is the most general case. Each task has its own set of evaluation measures which are explained in more detail in this section.

The performance statistics described in [394, 395, 396] are utilized to evaluate the system. Computing the performance of a biometric identification system generally requires three different image sets: The gallery $\mathcal{G}$, also referred to as the training set, and two probe sets, i.e. biometric samples which are presented to the system to report their identity. Let $\mathcal{P}_\mathcal{G}$ be the probe set that contains face images of primates in the gallery and $\mathcal{P}_\mathcal{N}$ the probe set that contains samples of individuals that are not known to the system. Note that although the images in $\mathcal{P}_\mathcal{G}$ contain individuals that are known to the system, the samples in the gallery $\mathcal{G}$ are different from the samples in $\mathcal{P}_\mathcal{G}$. When a probe p_j is presented to the system, a score vector $\boldsymbol{s} \in \mathbb{R}^{C\times 1}$ can be calculated, where C is the number of known individuals in the database. The entries of this vector are scaled between 0 and 1. Since confidence measures are used for classification within the proposed system, it can be stated that the higher the score of the predicted class, the higher the confidence of the classifier. For the proposed decision fusion technique described in Section 4.2.3.3, the combined weightings $\boldsymbol{s}_{\mathrm{df}}$ can be utilized as score values for each class.

Closed-set identification is the classic performance measure used in most scientific publications. For closed-set identification, the identity of the biometric test sample is always known to the system, i.e. it is assumed that every probe can be classified as one class represented in the training set. Open-set identification is the general case where the system first has to decide if the probe p_j represents a sample of an individual in the gallery $\mathcal{G}$ or not. If the system decides that the individual of the probe is genuine, then it also has to report the identity of the individual, otherwise it is classified as *unknown*.

5.3.1. Closed-Set Identification

When the number of individuals to be recognized is restricted, like in captive environments for instance, the closed-set identification task is the most common method to evaluate an identification system. However, instead of asking if only the top-match is correct, a more general approach would be to ask if the correct individual is among the top $\wp$ matches.

After computing the scores of the probe p_j belonging to all classes in the training set, the first step is to sort the resulting vector of score values and compute the rank of p_j. Since the classification of the PRF is based on confidence metrics, it has to be sorted descendingly so that the best guess of the system appears as the first element of the score vector. Similarly, if a system is based on distance metrics the vector has to be sorted ascendingly. Consequently, the $\wp$-th rank of the probe p_j is then given by the $\wp$-th element of the score vector $\boldsymbol{s}_{\mathrm{df}}$. The rank-$\wp$ identification rate, $P_I(\wp)$, is therefore given by

$$P_I(\wp) = \frac{|C(\wp)|}{|\mathcal{P}_\mathcal{G}|}, \tag{5.1}$$

where $C(\wp)$ is the cumulative count of the number of probes that are less than or equal to rank $\wp$.

$$C(\wp) = |\{p_j : rank(p_j) \leq \wp\}| \,. \tag{5.2}$$

The identification rate at rank 1 is also referred to as top match rate, accuracy, or correct identification rate. For closed-set identification, the performance of a biometric identification system is often reported as Cumulative Match Characteristic (CMC), which plots $P_I(\wp)$ as a function of rank $\wp$.

Furthermore, it is often useful to consider the correct identification rate of every class separately. An appropriate tool for that purpose is called the *confusion matrix*, which provides an overview of the error distribution of a classifier for a multi-class problem. Thus, the confusion matrix shows which classes are more often confused by the system than others. Given C classes, the confusion matrix has a size of $C \times C$. Each row of the matrix represents the actual class given by the ground-truth information, while each column represents the class predicted by the classifier. Figure 5.9 shows a schematic overview of a confusion matrix.

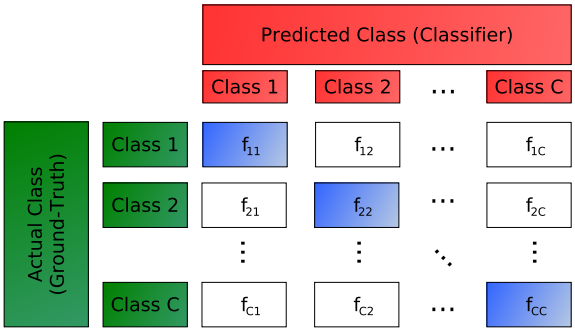


Figure 5.9.: ***Schematic overview of a confusion matrix.*** Each row of the square matrix represents the actual class given by the ground-truth information, while each column represents the class predicted by the classifier. Every element of the matrix contains the relative frequency $f_{i,j}$ that samples of class i were classified as class j. Therefore, the accuracy of the system can be determined by the mean of the elements of the main diagonal.

Every element of a confusion matrix (i, j) contains the relative frequency $f_{i,j}$ that samples of class i were classified as members of class j. Hence, all values that are not on the main diagonal of the confusion matrix correspond to misclassifications while the accuracy or rank-1 recognition rate of the system is given by

$$P_I(1) = \frac{\sum_{i=1}^{C} f_{i,i}}{C}. \tag{5.3}$$

5.3.2. Open-Set Identification

When it comes to evaluate not only the face recognition part but complete recognition systems including face detection, pre-processing, and identification (see Figure 4.1), a closed-set evaluation protocol is not applicable anymore. The reasons for this are twofold: First, the face detection stage will produce false positive detections which have to be considered for evaluation. Secondly, often not all individuals are known to the system in real-world applications. Hence, it is important that a biometric identification system is able to reject unknown individuals so they can be annotated by experts before being included into the training set. Besides manually annotating rejected data, a number of sophisticated algorithms for unsupervised clustering [397, 398] exists which could be applied to automatically group similar looking faces.

Therefore, an open-set identification scheme has to be utilized to evaluate systems that operate in real-world environments. While for a closed-set identification the question is how many test images are correctly classified as a certain individual, two more types of errors can occur for an open-set classification: In addition to false classifications, it is also possible that the system rejects genuine individuals or accepts impostors.

A probe p_j is detected and identified if the maximal match score $s_{max,j} = \max\{\boldsymbol{s}_{\mathrm{df},j}\}$ is above the operating threshold τ and identified correctly with $rank(p_j) = 1$. Therefore, the detection and identification rate P_{DI} at threshold τ can be calculated as

$$P_{DI}(\tau, 1) = \frac{|\{p_j : p_j \in \mathcal{P}_\mathcal{G}, rank(p_j) = 1, \text{and } s_{max,j} \geq \tau\}|}{|\mathcal{P}_\mathcal{G}|}. \tag{5.4}$$

The second performance measure is the false acceptance rate P_{FA}. A false acceptance occurs if the maximal match score of an impostor is above the operating threshold τ. Consequently, the false acceptance rate is the fraction of probes in $\mathcal{P}_\mathcal{N}$ that are detected as genuine individuals:

$$P_{FA}(\tau) = \frac{|\{p_j : p_j \in \mathcal{P}_\mathcal{N}, s_{max,j} \geq \tau\}|}{|\mathcal{P}_\mathcal{N}|}. \tag{5.5}$$

An ideal system would have a detection and identification rate P_{DI} of 1.0 and a false acceptance rate P_{FA} of 0.0, which means that all individuals are detected and classified correctly and no impostor has been erroneously classified as a genuine individual. In practice, however, both measures have to be traded-off against each other. This trade-off is shown in a Receiver Operating Characteristic (ROC) by iteratively changing the operating threshold τ.

Two important performance measures can be derived from the ROC-plot. The Equal Error Rate (EER) is reached when the false acceptance rate is equal to the false rejection and classification rate. The smaller the EER, the better performs the system. A second important performance measure is the Area under the ROC-curve (AUC), a measure between 0 and 1

where a perfect system would have an AUC of 1. Figure 5.10 exemplarily shows a ROC-curve and illustrates according performance measures.

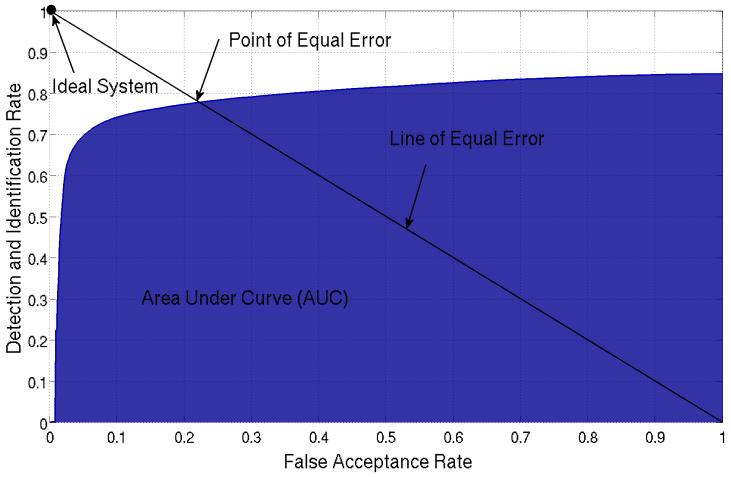


Figure 5.10.: ***Example of a ROC-curve.*** An ideal system would have a false acceptance rate P_{FA} of 0 and a detection and identification Rate P_{DI} of 1. The black line denotes the Line of Equal Error and the Area Under the Curve (AUC) is characterized in blue. Another important performance measure is the Equal Error Rate (EER) which is reached at the point where the false acceptance rate is equal to the false rejection and classification rate.

5.3.3. Experimental Design

Validation techniques are motivated by two fundamental problems in pattern recognition: model selection and performance estimation. Note that parameters are set to default values and therefore no parameter tuning is necessary for the proposed PRF. Therefore, the main question is how to obtain results which are a valid estimate of what can be expected in real-world settings.

To get a meaningful estimation of the performance of a classification system on a given dataset, it has to be split into disjoint subsets, the test and the training set. However, in a single train-and-test experiment on a small dataset the estimate of the error rate might be misleading due to an "unfortunate" split. This limitation can be overcome by applying resampling methods such as the K-fold cross-validation with the expense of a longer evaluation time.

In a K-fold cross-validation, the dataset is randomly divided into K subsets. For each of the K experiments $K-1$ folds are put together to form the training set and the remaining fold is used for testing. This procedure is repeated K times until every fold has been used as the test set once. The performance of the system is then estimated by computing the average accuracy across all K trials. Figure 5.11 illustrates the procedure of a K-fold cross-validation exemplarily on a dataset of two classes and 30 samples.

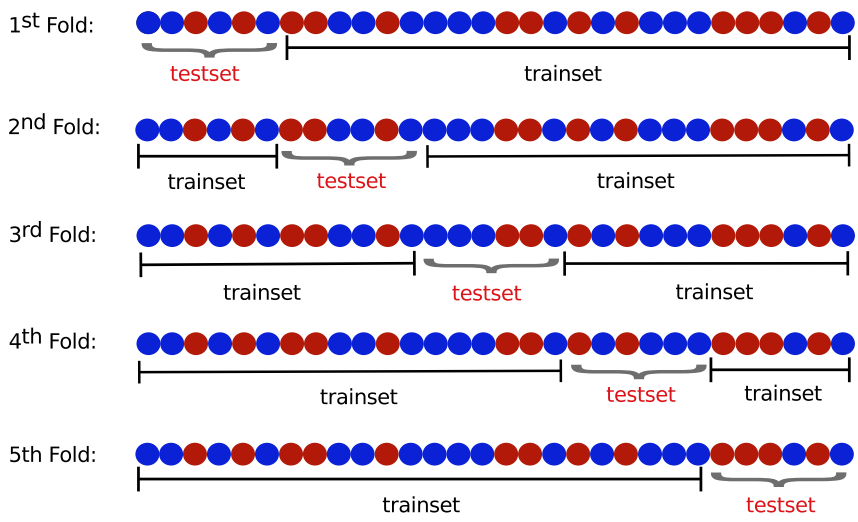


Figure 5.11.: ***Illustration of a K-fold cross-validation.*** The figure illustrates the procedure of a K-fold cross-validation on a dataset of two classes (blue and red dots) and 30 samples using 5 folds. The dataset is split into K subsets. For each of the K iterations, the samples of $K-1$ subsets form the training set while the remaining one takes the role of the test set. This ensures that each sample in the dataset gets to be a test sample exactly once. Unless not otherwise specified, a 5-fold cross-validation was applied for the experiments within this thesis to ensure validity of the performance estimation.

A disadvantage of a K-fold cross-validation is that the split into training and test set might be unbalanced, e.g. a certain class is under-represented in the training or test set. Therefore, an extension of the ordinary K-fold cross-validation is its stratified version where it is taken into account that each class is approximately equally represented in both sets.

Unless otherwise specified, a stratified 5-fold cross-validation was applied for all conducted experiments to obtain a valid estimation of the performance of the proposed system. Additionally, for open-set experiments one individual at a time is removed from the training set and presented as an impostor to test the system's capability to reject unknown individuals. This procedure is repeated C times, where C is the number of classes in the data set, such that every individual takes the role of an impostor once. Furthermore, the detection stage of the proposed primate recognition framework might produce false positive detections in real-world settings which are subsequently processed by the facial recognition module. To test the system's performance to reject non-face data, false positive detections are assigned to the *unknown* class. Images of false positive detections as well as all pictures of the unknown individual remain in the test set for all 5 folds and are not used for training.

5.4. Individual Identification in Images

5.4.1. Identification Using Global Features

During the course of the subsequent sections the proposed approach using global features (see Section 4.2.3.1) is benchmarked against eight different state-of-the-art face recognition techniques developed for human face recognition over the past 20 years. All algorithms were discussed in Section 3.3 and are briefly reviewed below as a reminder.

Pixel-based Algorithms: The algorithms *Eigenfaces (PCA)* [61], *Fisherfaces (LDA)* [62], and *Laplacianfaces (LPP)* [58] use simple gray-level pixel information as features. They only differ in the applied feature space transformation technique. A simple Euclidean distance-based k-Nearest-Neighbor (k-NN) classifier was used for all three algorithms with $k = 3$ as proposed in the original publications. *Randomfaces (RAND)* [71] on the other hand uses a randomly generated projection matrix by sampling zero-mean independent identically distributed Gaussian entries for dimensionality reduction. Furthermore, the Sparse Representation Classification (SRC) scheme introduced in Section 2.4.2 is applied after feature space transformation. The algorithms *Robust Sparse Coding (RSC)* and *Regularized Robust Coding (RRC)*[3] utilize pixel-based information for features representation as well. However, both algorithms apply different extensions of SRC for classification which were proposed in [313] and [314], respectively. Note that for RSC a PCA is applied to project the high-dimensional vectorized images into a smaller dimensional subspace. For RRC, however, a feature space transformation is not required.

Advanced Feature-based Algorithms: Moreover, algorithms that compute more advanced feature sets were implemented for benchmark purposes. The algorithm developed by Ekenel and Stallkamp in [315, 316, 307] applies a block-based Discrete Cosine Transform (BDCT) to obtain representative descriptors for each face. As proposed by the authors, each facial image is divided into 8×8 blocks and 10 DCT coefficients are extracted in each block by utilizing a zig-zag scan. The first element is ignored in order to compensate illumination variation. A feature space transformation step is not performed. As for *Eigenfaces*, *Fisherfaces*, and *Laplacianfaces*, a 3-NN approach is used for classification. However, in [307] the authors proposed to use the correlation between feature vectors as distance measure. Finally, the *Gabor Sparse Representation Classification (GSRC)* approach by Yang *et al.* proposed in [48] was implemented for comparison. As the name suggests, Gabor features are extracted and SRC is used for classification. A PCA was applied for feature space transformation as suggested by the authors.

[3] The authors provide the source codes of RSC and RRC at
`http://www4.comp.polyu.edu.hk/~cslzhang/papers.htm` Last visit: May 22^{nd}, 2014

For all algorithms the parameters were set as proposed in the original publications if declared by the authors. As suggested in Section 4.2.3.1, the dimensionality of the feature vector was set to $m = 160$ for all algorithms where feature space transformation needs to be performed in order to keep the comparison fair. Note, however, that for the *Fisherfaces* approach the maximum number of features is limited to $C - 1$, where C is the number of individuals in the dataset (see Section 2.3.2 for details). Thus, m is set to 23 and 48 for the ChimpZoo and the ChimpTaï dataset, respectively. Unless otherwise noted, a 5-fold cross-validation (see Section 5.3.3) was implemented using closed-set or open-set identification protocols in order to obtain valid results. All experiments in this section were conducted on the frontal face datasets listed in Table 5.3.

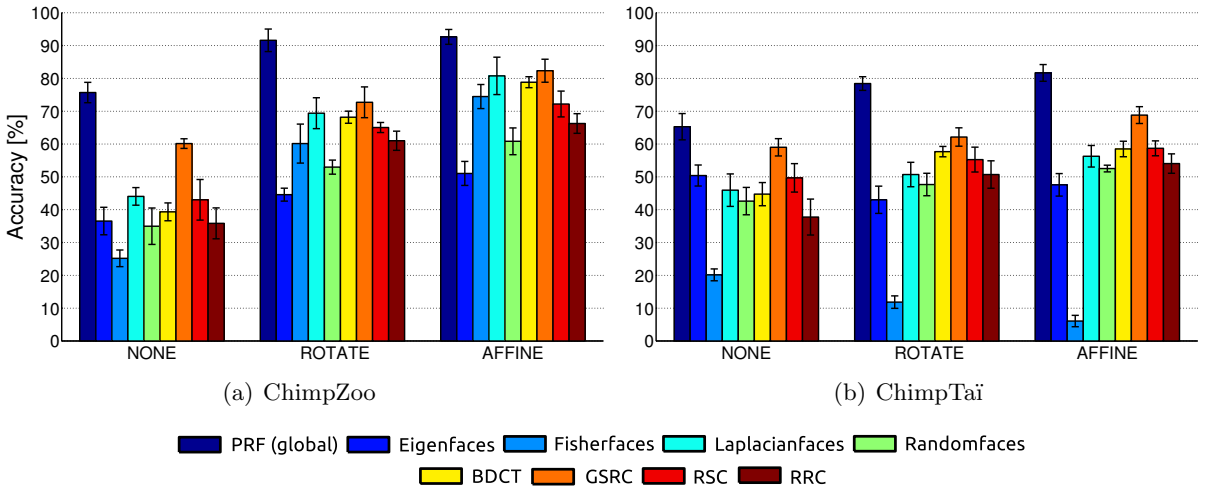
The Influence of Face Alignment

As stated in section 4.2.2, face alignment is an important pre-processing step to achieve high accuracies in real-world settings. Hence, the goal of the first experiment is to compare the influence of different alignment techniques applied to the state-of-the art face recognition techniques outlined above as well as the global face recognition pipeline (see Section 4.2.3.1) as part of the proposed PRF. Three different alignment strategies are compared:

- **NONE**: No alignment is applied. An example of an unaligned picture from the ChimpZoo dataset is illustrated in Figure 4.4(a).
- **ROTATE**: The face is rotated to an upright position such that both eyes lie on a horizontal line. A rotated chimpanzee face can be seen in Figure 4.4(b). Based on the re-located eye and mouth markers the image is subsequently cropped to further reduce undesirable background.
- **AFFINE**: The proposed face alignment scheme discussed in Section 4.2.2 is applied which performs an affine transformation based on the locations of both eyes and the mouth. This ensures that facial feature points are located at predefined position throughout the entire dataset. A transformed chimpanzee face from the ChimpZoo dataset can be found in Figure 4.4(c).

Each face is subsequently resized to 64×64 pixels. Only full-frontal face images were taken into account that met the requirements discussed in Section 5.2.3 for the experiments conducted in this section.

Figure 5.12(a) shows the results obtained for the ChimpZoo dataset while the rank-1 accuracies as well as the standard deviations across the folds for the ChimpTaï dataset can be seen in Figure 5.12(b). A more detailed overview of the achieved results can be found in Tables D.1 and D.2.



(a) ChimpZoo (b) ChimpTaï

Figure 5.12.: ***Results of state-of-the-art algorithms for three alignment strategies***. The figure shows how different state-of-the-art face recognition algorithms perform when three different alignment strategies are applied: *No alignment (NONE)*, *rotation (ROTATE)*, and *affine transform (AFFINE)*. Accuracies are illustrated in bar plots for the ChimpZoo dataset (a) and the ChimpTaï dataset (b). Additionally, the standard deviation across the 5 folds is superimposed for each bar. The legend is plotted below the diagrams. It is obvious that the proposed PRF global face recognition pipeline (dark blue) outperforms all other approaches for all three alignment strategies, where the mean accuracy for AFFINE is slightly higher than for alignment based on rotation.

It is obvious that the global face recognition pipeline of the proposed PRF, represented by the dark blue bars, outperforms the other approaches on both datasets, regardless of the applied face alignment strategy. As expected, the mean accuracies for ChimpTaï are lower than for the ChimpZoo dataset for all face recognition algorithms. As already discussed in Section 5.2.2, the ChimpTaï dataset was gathered under more challenging conditions in natural habitats of free-living chimpanzees. Additionally, it contains more individuals which explains the worse accuracies obtained for this dataset. Note, however, that a proper alignment increases the performance of most face recognition algorithms which proofs the statement that face alignment is crucial to achieve high accuracies in real-world settings. The experiments furthermore show that an affine transformation based on three distinctive keypoints is very well suited as face alignment step since it performed best for most face recognition algorithms. Accuracies increase up to 15% on both datasets for the proposed framework. Another interesting observation is

that for the ChimpTaï dataset the simple *Eigenfaces* approach seems to be superior to the more advanced *Fisherfaces* algorithm which takes class information into account. This phenomenon has already been discussed in Section 2.3.2 and in [60]. It is well known that PCA can outperform LDA if the the underlying class distribution cannot be adequately represented by the training data which might particularly be the case for the ChimpTaï dataset.

One can also see from the results in Figure 5.12 that representative image descriptors as well as a suitable feature space transformation is necessary to obtain high accuracies. For pixel-based algorithms, a sophisticated feature space transformation such as LPP outperforms methods which are based on simpler methods such as PCA or LDA. This supports the assumption that face images reside on a non-linear submanifold hidden in the feature space. By taking the local context between images of the same class into account, adequate performances can be achieved even with pure gray-value information as features. However, *GSRC* and *BDCT* also perform very well for both datasets suggesting that a compact and robust representation of a face is equally important. Thus, the combination of the face descriptor, proposed in Section 4.2.3.1, and state-of-the-art feature space transformation and classification techniques provides a robust and accurate face recognition system for identification of great apes in real-world settings.

The Influence of Image Enhancement Techniques

As claimed by Moses *et al.* in [349], the variations of face images of the same individual caused by various environmental factors are often larger than image differences caused by the identity itself. Therefore, many algorithms have been proposed in the literature to compensate illumination changes and counter the effects of local shadowing in order to help increasing robustness of face recognition algorithms.

Hence, the second experiment investigates how several state-of-the-art photometric normalization techniques influence the recognition accuracies of face recognition algorithms. The proposed global feature approach of PRF is benchmarked against the same face recognition techniques outlined above. Since face alignment based on an affine transform achieved the best results for most algorithms as shown in Section 5.4.1, this pre-processing step is applied for all face recognition methods in this experiment. The following illumination normalization techniques are compared:

- **NONE**: Except for face alignment, no additional pre-processing steps are applied.
- **HIST**: A histogram equalization is performed after face alignment which enhances the contrast of an image by transforming the values of an intensity image such that the gray-level histogram of the output image is approximately equally distributed [399].

- **CLAHE**: A Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied after alignment to compensate illumination changes. Opposed to an ordinary histogram equalization, CLAHE operates on small image regions. A histogram equalization is performed separately for every region in order to enhance local details. To compensate noise effects in homogeneous regions, CLAHE additionally applies a parameter that specifies a contrast enhancement limit. Neighboring regions are then combined using bilinear interpolation to eliminate artificially induced boundary effects [400].

- **MSR**: The Multi Scale Retinex (MSR) algorithm is a photometric normalization technique proposed by Jobson *et al.* in [401]. It is based on the so-called *retinex theory* which was introduced in [402]. The implementation of MSR was taken from the *INFace toolbox*[4] [403], a collection of various photometric normalization techniques.

- **IMADJUST**: An image adjustment algorithm is applied after face alignment to enhance the contrast of an image. The algorithm maps the intensity values of a grayscale image such that 1% of the data is saturated at low and high intensities of image $\boldsymbol{I}$. IMADJUST is part of the MATLAB® Image Processing Toolbox.

- **GAMMA+DOG**: A chain of preprocessing steps for illumination normalization and contrast enhancement proposed by [40] is applied. The method includes a series of steps. First, a non-linear gray-scale transformation called gamma correction is performed. The intensity values of image $\boldsymbol{I}$ are replaced with $\boldsymbol{I}^{\gamma}$ if $\gamma > 0$ or $\log(\boldsymbol{I})$ for $\gamma = 0$, where $\gamma \in [0, 1]$ is a user-defined parameter. This procedure enhances the local dynamic range in dark or shadowed regions while compressing it in bright regions. For the experiments in this thesis, γ is set to 0.2. Gamma correction alone does not completely remove the influence of shading effects caused by low frequency components of illumination gradients. However, suppressing high frequency components might also help to increase the recognition performance since noise and aliasing effects can be removed. Thus, Difference of Gaussian (DoG) filtering is performed after gamma correction in order to obtain a bandpass filtered image. The input image is filtered with two 2D Gaussian kernels of different standard deviations and the output images are subsequently subtracted to obtain the bandpass filtered image. As suggested in [40], $\sigma_1 = 1.0$ and $\sigma_2 = 2.0$ were set as default setting for the experiments conducted below.

Figure 5.13 shows the effect of the applied photometric normalization techniques as well as the according histograms of the normalized images.

[4] Available for download at
`http://luks.fe.uni-lj.si/sl/osebje/vitomir/face_tools/INFace/` Last visit: May 20^{th}, 2014

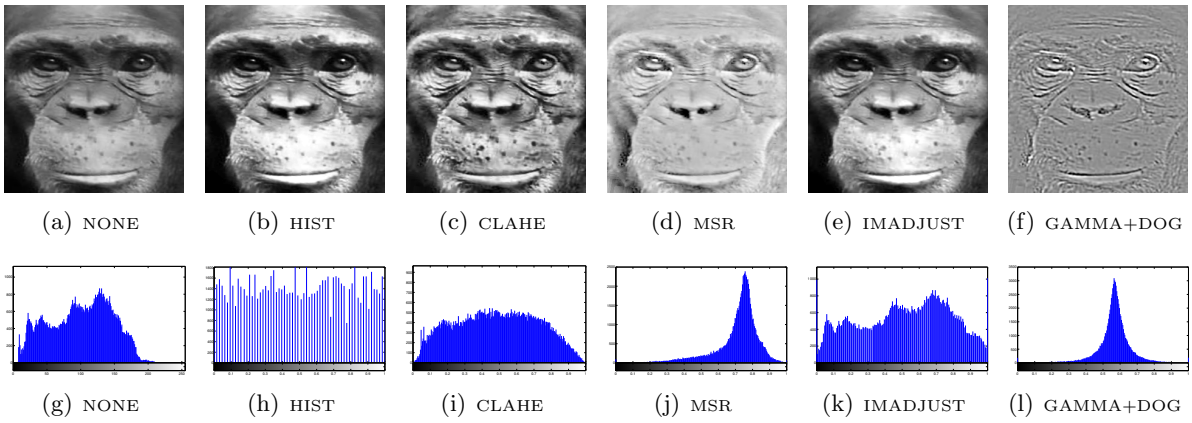


Figure 5.13.: ***The effect of different illumination normalization methods.*** The figure illustrates the effect of the five different photometric illumination normalization techniques used for experimentation in this thesis. The methods were applied to an aligned chimpanzee face (a)-(f). The according gray-level histograms of the processed images are illustrated below.

The obtained results of the investigated face recognition algorithms in combination with the lighting normalization techniques explained above can be seen in Figure 5.14. For illustration purposes the standard deviations across the folds are not displayed. However, the results for both datasets including the standard deviations are given in Tables E.1 and E.2.

The obtained results show that illumination normalization techniques do not have a significant impact on the proposed global face recognition pipeline. Hence, the consistent accuracy for both datasets indicate that the proposed global face descriptor is to some extent invariant to challenging lighting conditions compared to pure pixel-based information. An additional pre-processing step for compensation of lighting variations is therefore not necessary since none of the applied pre-processing methods has a significant effect on the recognition accuracy. However, it is obvious that lighting normalization can increase the performance of face recognition algorithms which solely rely on gray-scale pixel intensities quite drastically. Surprisingly, simpler methods such as HIST or CLAHE often perform better than more advanced algorithms like MSR for instance. Another interesting observation is that Gabor features alone seem to be prone to challenging lighting situations since HIST and CLAHE are able to increase the accuracy of GSRC for both datasets. The proposed ELGTPHS descriptor on the other hand is based on Gabor wavelets as well. However, additionally extracting ELTP features seem to be able to increase the robustness of the overall descriptor against illumination changes which is a prerequisite in real-world environments. Also the BDCT descriptor proposed by Ekenel *et al.* in [316, 307] does not benefit from an additional lighting normalization step.

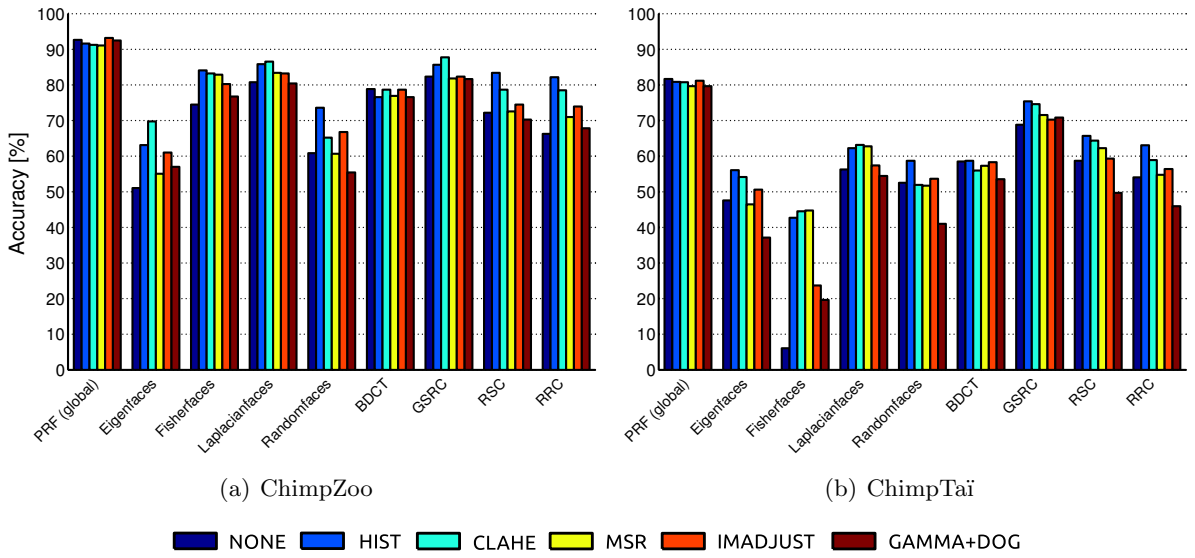


Figure 5.14.: ***The influence of different illumination normalization methods.*** The figure shows the influence of different preprocessing techniques for illumination normalization for the ChimpZoo dataset (a) and the ChimpTaï dataset (b). A legend is plotted below the diagrams. The results show that an additional lighting normalization step has almost no effect on the recognition results of the proposed identification algorithm, whereas illumination normalization can significantly improve the performance of algorithms which are solely based on gray-scale pixel information.

This is not surprising since the first element of any DCT transformed block (the DC component) is omitted during feature extraction. Extracting the AC coefficients of the DCT transformed signal and subsequently normalizing the feature vector to unit norm already ensures the illumination invariance of the descriptor.

Evaluation and Enhancement of Face Recognition Algorithms using Synthesized Data

Previously, different state-of-the-art face recognition algorithms and preprocessing steps were evaluated on full-frontal face images. The uncontrollable nature of great apes as well as the challenging environment in which visual footage is usually captured by biologists require face recognition algorithms to be robust to pose and illumination variance. Thus, this section explores face recognition accuracy on face images with variations of pose and lighting. However, thoroughly profiling face recognition algorithms with data collected under natural conditions often is cumbersome since

(a) a huge amount of extensively annotated data is required which is hard to achieve in real-world environments, and

(b) objectively annotating parameters such as pose angles, illumination intensities, and lighting direction for data gathered under uncontrolled conditions is hardly feasible.

Moreover, the elusive nature of great apes often limits availability of data with sufficient variety of image parameters. Thus, decent training of face recognition algorithms to cope with extreme lighting conditions and different head poses is often not possible. Moreover, as stated in Section 5.2.3, a minimum amount of images per individual is required in order to perform proper training which might result in individuals being omitted from analysis due to insufficient data. By synthesizing 2D images from a custom-built generic 3D model with controlled pose and illumination variations, the above mentioned challenges might be avoided. A generic 3D model offers the possibility to synthesize face images of individuals in various controlled poses and lighting conditions. Hence, the robustness of face recognition algorithms against multiple challenges present in real-world environments can be evaluated thoroughly. Furthermore, increased robustness of face recognition algorithms against pose and illumination variation might be achieved without requiring more real data but instead augmenting training data with automatically generated synthesized data. Such an approach might not only save manual effort but also reduce the amount of animals which has to be discarded due to insufficient data. Therefore, the objectives of this section are twofold:

1. To quantify the relationship between pose and illumination parameters and algorithm accuracy. Automatically generated synthesized test data with different controlled image variations are used to highlight limitations of selected face recognition algorithms and identify operation constraints when trained on real full-frontal data.

2. To increase the generalization capability of the algorithms for horizontal pose variations. Synthetic face images rendered with different poses are used to supplement the training data in order to obtain a higher robustness against off-frontal poses.

Details about the dataset used for the subsequent experiments were given in Section 5.2.4. Synthetic images were created from the ChimpZoo dataset only by altering the least complex parameters, pose and illumination, in order to explore their effects on face recognition algorithms. As already discussed in Section 5.2.4, pose variation has been split into two categories: horizontal and vertical pose variation each varied between $\pm 30^\circ$ in 10° increments. The lighting conditions on the other hand contain an ambient setting and can include a spotlight. Its position is varied between $\pm 60^\circ$ in 30° increments, while its intensity contains three levels: high, mid, or low. Only one parameter is varied at a time while others remain in a neutral state of ambient lighting or frontal pose, respectively. For each of the subsequent experiments again a 5-fold cross-validation was applied to obtain valid results.

To guarantee separability between training and test sets, each synthetic image is associated with a *"seed"* image. Hence, rendered images only ever appear in the train or test set in which the seed image is present. Experiments were conducted in collaboration with the *Visual Information Laboratory, University of Bristol, UK,* and were partially published in [10]. The five best performing algorithms from the previous sections were selected for evaluation: *PRF (global)*, *GSRC*, *Laplacianfaces*, *RSC*, and *RRC*. The parameters of each algorithm were set as discussed in the previous sections.

Profiling Face Recognition Algorithms Using Synthetic Test Data

Experiment 1: Real images for training, synthetic images for testing:
The objective of the first experiment is to quantify the influence of pose and illumination on the accuracy of face recognition algorithms. Therefore, every algorithm is trained with real full-frontal images from the ChimpZoo dataset and synthetically rendered data is used for testing. Experiments were conducted on four synthetic datasets: one each for horizontal and vertical pose variation, lighting exposure, as well as spotlight position. Every real frontal test image is replaced with its synthetically generated counterparts for each variation.

Experiment 2: Synthetic images for training, synthetic images for testing:
The goal of the second experiment is to increase the system's robustness to pose and lighting variation. Each system is trained with synthetic image sets where only one parameter varies throughout its entire range. For instance if the objective is to increase robustness to horizontal pose variation, face recognition algorithms are trained on sets of synthetic images in each horizontal pose in place of the real training image. The trained models are then tested on their according synthetic test sets. It is guaranteed that training and test sets are completely separated for every fold. The motivation is that a richer training set should provide a more dense representation of the image variability and thus increase the system's generalization capability.

The results for both experiments can be found in Figure 5.15. The black dashed lines with circle markers denote the results obtained for the first experiment (real training data, synthetic test data). The accuracies for the second experiment (synthetic training data, synthetic test data) are denoted with blue dashed lines. Details about the obtained results including the standard deviations across the five folds can be found in Tables F.1 - F.3.

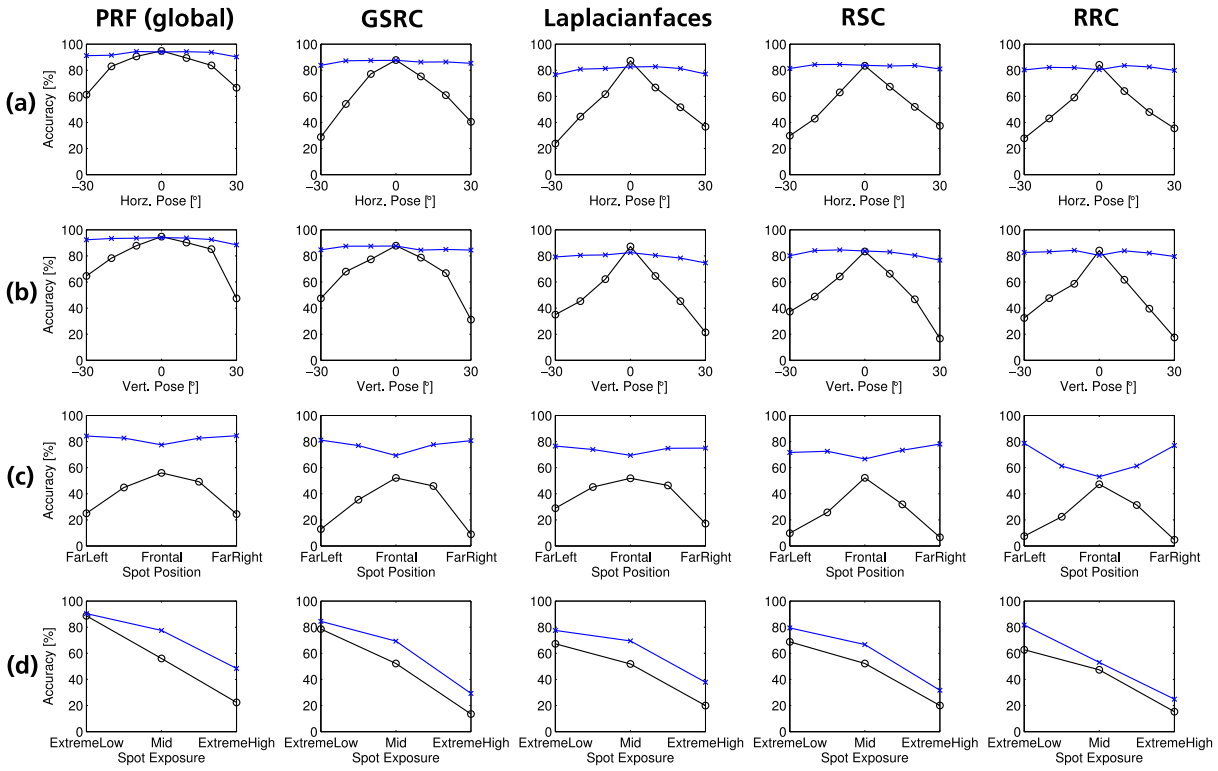


Figure 5.15.: ***Face recognition accuracies for training and testing with synthesized data.*** The figure shows the obtained accuracies for selected face recognition algorithms over horizontal (a) and vertical (b) pose variations as well as different spotlight positions (c) and exposures (d). Black lines with circle markers are results obtained when the system was trained on real full-frontal face images and tested on synthesized data. Blue lines with crosses denote the accuracies if the system was trained on relevant synthesized data. If the algorithms are trained on relevant artificial data, the accuracies of all systems stabilize for off-frontal data and images with extreme lighting conditions, respectively.

Influence of pose variation: It can be seen from the black graphs in Figure 5.15 (a) and (b) that the accuracy of all algorithms declines significantly when confronted with off-frontal posed faces, indicating a low generalization capacity beyond facial poses represented in the training set. The higher accuracy of PRF (global) for both vertical and horizontal pose variation compared to other algorithms indicates a more flexible representation of individuality by the proposed ELGTPHS descriptor compared to basic Gabor or pixel-based features. It can further be noted that for vertical pose variation the accuracies for all algorithms are asymmetrically distributed. For positive angles (face tilting upwards) the accuracy curves of all algorithms declining more significantly than for negative angles (face tilting downwards). This behavior is interpreted as greater invariance of all algorithms to downward poses than upward poses which might be due to more useful visual information of the upper half of a face.

This actually makes sense since the mouth region is often subject to severe occlusion and facial expressions caused by eating or other behavioral activities.

As expected, when synthetic training images are included in the training set, face recognition algorithms produce an almost uniform accuracy curve across pose variations which is denoted by the blue graphs in Figure 5.15. The accuracy for PRF (global) for instance is approximately maintained at 90% accuracy across the whole range of vertical and horizontal pose variations. This indicates that when algorithms are trained with data which sufficiently reflects the variance of the test set, all face recognition algorithms have the ability to generalize and therefore achieve better results compared to systems solely trained with full-frontal face images. It is also worth mentioning that the peak accuracy for full-frontal face images is maintained for almost all algorithms. This behavior suggests that additional data does not hamper performance although additional "noisy" training data increases the intra-class variance making separation of classes more challenging.

Influence of illumination variation: The accuracy plots for illumination variation can be found in Figure 5.15 (c) and (d). If trained on real full-frontal, well-lit images, all algorithms suffer from declining accuracy for different spotlight positions at mid-level intensity (black-line plots in Figure 5.15 (c)). The highest accuracy is obtained for a spotlight in full-frontal position and decreases for more extreme angles, suggesting that shading effects caused by non-frontal spotlight positions cannot be compensated well. As can be seen in Figure 5.7, spotlight positions in extreme angles introduce shadowed blacked-out patches at the opposite side of the face from the spotlight which can be considered as severe occlusion. While PRF (global) still preserves an accuracy of above 20% for spotlight angles of $\pm 30^{\circ}$, the recognition rates of RSC and RRC decrease to below 10%, suggesting that the proposed algorithm can handle occlusions better than competing face recognition methods. Furthermore, the intensity of lighting adversely affects the accuracy of all five algorithms (see black curves in Figure 5.15 (d)). While for low spotlight intensity at frontal spotlight position almost all algorithms obtain a similar recognition rate as for full-frontal ambient lit images, accuracy declines drastically for higher exposures. This is likely due to a white-out effect of facial features as visible in Figure 5.7. For high exposures, important features such as edges and wrinkles become invisible, making it hard for face recognition algorithms to distinguish between individuals.

Again, if synthetic data is included in the training set all algorithms increase robustness against the according variation (see blue curves in Figure 5.15 (c) and (d)). Especially for spotlight positions the accuracy of all algorithms is improved significantly. Interestingly, for more extreme angles the accuracy of all systems improve with respect to a frontal spotlight position, indicating that occlusion effects by strong shadows can be better compensated than

over exposure of the whole face if according training data is provided. This assumption is supported by the results for different lighting intensities. There still seems to be a sensitivity to strong lighting which is illustrated by the blue curves in Figure 5.15 (d). Although the overall accuracy for different spotlight exposures is increased for all algorithms, the accuracy still declines significantly for extreme intensities.

In summary, it seems to be beneficial to learn from additional data which introduce more extreme parameter variation rather than frontal, well lit face images alone. As can be deduced from the previous two experiments, training face recognition algorithms with representative data helps the system to generalize better and hence improves the system's robustness against challenging image variations. However, as yet, face recognition algorithms were trained and tested solely on synthetic data generated from a custom-built generic 3D model. Thus, the next two experiments investigate if this artificial data can be used to improve the robustness of face recognition algorithms when tested on real off-frontal data.

Improving Robustness by Augmenting Training Data with Rendered Images

Experiment 3: Real images for training, real images for testing:
In order to address the question if synthetically generated training data is suitable for training face recognition algorithms, first all five systems are trained on real full-frontal data and tested on off-frontal posed data to measure their baseline performances. Example images as well as detailed descriptions of the non-frontal subsets were given in Section 5.2.3.

Experiment 4: Synthetic images for training, real images for testing:
Subsequently, algorithms are trained on sets augmented with pose-offset synthetic data and tested on real off-frontal data as before. To guarantee a fair comparison of both methodologies, the cross validation sets are maintained for the experiments conducted in this section. For every real training image in the set four synthetic images are added, one each with $\pm 20^\circ$ horizontal and $\pm 20^\circ$ vertical pose offset. Hence, the resulting training sets contain only 20% real data and 80% synthetically generated images. Although face recognition algorithms are solely tested on horizontal offset-posed data, vertical and horizontal varied images are included in the synthetically augmented training sets to address the moderate tilt variation permitted in the test data.

Figure 5.16 visualizes the obtained results. Again, the black graphs with circle markers denote the accuracies obtained for algorithms solely trained with real images while the blue-lined plots denote the results if synthetically augmented training data was used. Accuracies as well as standard deviations for all algorithms across different poses are further given in Table 5.6.

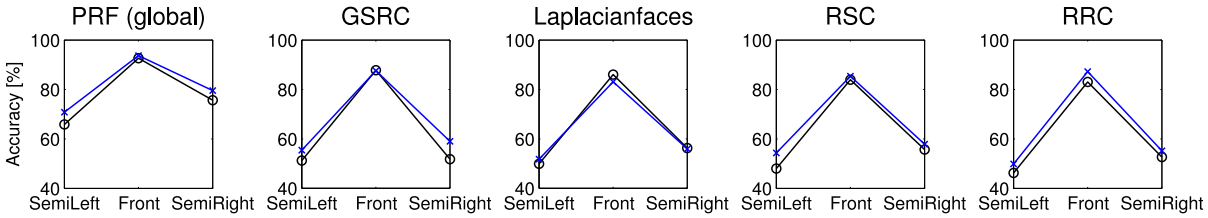


Figure 5.16.: ***Face recognition accuracies for algorithms trained on synthesized data and tested on real offset-posed data.*** Five face recognition algorithms were first trained on real full-frontal face images and tested on semi-left and semi-right face data (black graphs with circle markers). The same algorithms were then trained with synthesized offset-posed data and tested on the same real non-frontal face images as before (blue-lined plots with crosses). As can be seen, synthesized face images rendered from a generic 3D model can help to generalize face recognition algorithms and therefore increase their robustness to non-frontal face images.

Acc (Std.) [%]	**Train Data**	**SemiLeft**	**Front**	**SemiRight**
PRF (global)	real	65.86 (1.55)	92.64 (3.50)	75.64 (1.18)
	synthetic	**70.83 (1.17)**	**93.69 (2.87)**	**79.56 (0.82)**
GSRC	real	51.23 (0.68)	**87.80 (4.59)**	51.79 (1.52)
	synthetic	**55.41 (0.55)**	87.59 (3.74)	**59.01 (1.17)**
Laplacianfaces	real	49.94 (1.25)	**86.05 (2.48)**	**56.31 (1.69)**
	synthetic	**51.77 (1.33)**	83.09 (3.42)	55.92 (1.24)
RSC	real	47.99 (1.16)	83.97 (6.40)	55.69 (1.34)
	synthetic	**54.31 (1.31)**	**85.35 (5.89)**	**57.88 (1.23)**
RRC	real	46.22 (1.05)	83.03 (5.01)	52.68 (1.12)
	synthetic	**49.90 (1.26)**	**87.26 (3.37)**	**55.18 (0.75)**

Table 5.6.: ***Accuracies and standard deviations for algorithms trained on synthesized data and tested on real off-frontal data.*** The table compares the accuracies and standard deviations of five face recognition algorithms tested on both full-frontal and off-frontal posed face images. Models were either trained on real frontal face data or synthetically generated data containing semi-profile face images. The best accuracies are printed in boldface letters. Improvements for all algorithms except *Laplacianfaces* can be achieved for off-frontal poses if synthetically generated data is included in the training set while the peak accuracies for full-frontal face images are maintained.

The obtained results indicate that synthesized training data rendered from a generic 3D model is indeed suited to increase to robustness of face recognition algorithms against off-frontal posed face images gathered under real-world conditions. Except for the *Laplacianfaces* approach, the accuracy of all algorithms is at least maintained for full-frontal face images and even increased for off-frontal posed face data. Hence, including artificially rendered images into the training set does not harm the algorithm's performance for full-frontal face data but instead helps to improve generalization capacity to better cope with slight pose variations. Although it could not be evaluated thoroughly due to lack of sufficiently annotated test data, same improvements might be expected for images gathered in varying lighting conditions.

Despite the obtained improvements, the increase in performance for off-frontal real data is rather marginal compared to the results obtained in the previous experiments, where algorithms were trained and tested on synthetic data. This is likely due to the lack of complexity of the used 3D model, suggesting that the model is capable of imitating parameter variations present in real-world environments only to a certain extend. The development of more sophisticated 3D models of chimpanzee faces might therefore be one key step to further increase the robustness of face recognition algorithms. Once such a generic face model has been developed, more realistic data could be rendered which might even lead to *"one-shot learning"*: Only a single real face image might be needed to generate enough synthetic training data to adequately train face recognition algorithms in order to perform well even under real-world conditions.

5.4.2. Identification Using Local Features

After thorough evaluation of the proposed face recognition pipeline using solely global features, the performance of different state-of-the-art local keypoint descriptors is investigated in this section. For this purpose, three different local interest point descriptors are compared. Most local features comprise two steps: Keypoint detection and description. However, for the application in this thesis only the descriptor is utilized and evaluated. Furthermore, often keypoint descriptors are designed to be rotation invariant which usually is achieved by applying a preprocessing step which detects the main gradient orientation of the interest point. However, the proposed face recognition pipeline already includes a face alignment step which makes the detection of the main orientation expendable. Therefore, all local keypoint descriptors are extracted without the property of rotation invariance for the sake of computational performance. The following keypoint descriptors are compared against each other in this section:

- **SIFT:** Scale Invariant Feature Transform (SIFT)[5] was presented by David Lowe in 2004 [51] and was briefly reviewed in Section 2.2.2.
- **SURF:** Speeded-Up Robust Features (SURF)[6] was developed by Bay *et al.* in 2008 [52] as an extension of SIFT. It is claimed by the authors that SURF is several times faster, more compact and at the same time more robust against certain image transformations than SIFT. An introduction to SURF can be found in Section 2.2.2. For the experiments conducted in this section, the upright version of SURF is used since it is claimed by [52] that U-SURF is faster to compute and can even improve the performance of the system.

[5]The VLFeat open source library (version 0.0.17) was used to extract SIFT descriptors. `http://www.vlfeat.org/` Last visit: May 23^{rd}, 2014

[6]OpenSURF was used in this thesis for SURF feature extraction. `http://www.chrisevansdev.com/computer-vision-opensurf.html` Last visit: May 23^{rd}, 2014

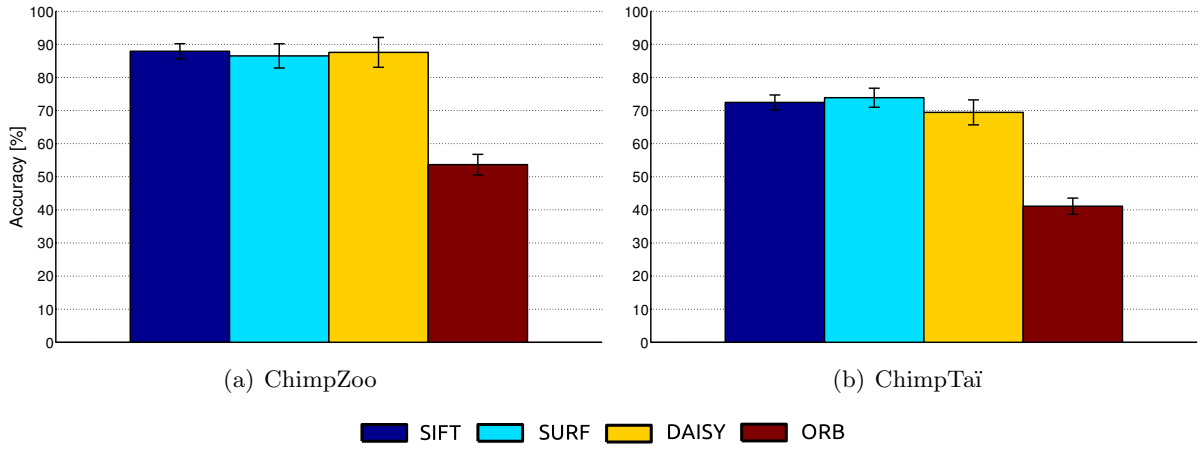
- **DAISY:** Recently, an efficient local descriptor named DAISY was proposed by Tola *et al.* in [404]. DAISY[7] was originally developed for wide-baseline stereo matching. As many keypoint descriptors, DAISY is inspired from earlier gradient-based interest point detectors and descriptors such as SIFT. However, it can be computed more efficiently and is better suited for depth estimation and occlusion detection from stereo images. For a detailed description of DAISY the reader is referred to [404]. All parameters, except for the radius of the descriptor, were set to the default values as given in Table 1 of the original publication [404].

- **ORB:** Oriented FAST and Rotated BRIEF (ORB)[8] was recently proposed as efficient alternative to SIFT and SURF by Rublee *et al.* in [230]. ORB is a substantial extension of Binary Robust Independent Elementary Features (BRIEF) [405], a fast binary descriptor. It is claimed by Rublee *et al.* that ORB is rotation invariant, resistant to noise and about two orders of magnitude faster than SIFT. For a detailed explanation of ORB the reader is referred to [230].

As described in Section 4.2.3.2, local keypoint descriptors are extracted around six distinctive interest points. The locations of the facial landmarks are determined based on the eye and mouth markers. Figure 4.9(b) shows a chimpanzee face and the superimposed keypoints for local feature extraction. Again, an affine transformation is applied for face alignment. Additional pre-processing steps for lighting normalization or contrast enhancement were not applied. After feature extraction, the obtained vectors are concatenated to form a single descriptor. To further reduce its size, LPP is used to transform the resulting high-dimensional feature vectors into a lower dimensional subspace of size $m = 160$. For classification, an SVM is trained using a 3-step grid search with exponentially growing values of C and γ. More details about the parameter settings can be found in Table C.1. In order to obtain valid results, again a 5-fold cross-validation (see section 5.3.3) was implemented using a closed-set identification protocol. The experiments in this section were conducted on the frontal face datasets for which details are given in Table 5.3. The obtained results are illustrated in Figure 5.17 and Table 5.7.

[7] The authors provide an implementation of DAISY which can be found at `http://cvlab.epfl.ch/software/daisy` Last visit: May 23^{rd}, 2014

[8] ORB is part of the freely available OpenCV library (version 2.3) which can be downloaded at `http://opencv.org/downloads.html` Last visit: May 23^{rd}, 2014



(a) ChimpZoo (b) ChimpTaï

Figure 5.17.: ***Results of the face recognition pipeline using local features.*** The figure shows the accuracies and standard deviations of the proposed local face recognition pipeline for different keypoint descriptors. Again, the experiments were independently conducted for the two datasets (a) ChimpZoo and (b) ChimpTaï. A legend is plotted below the diagrams. Four state-of-the-art local keypoint descriptors were compared. The results of SIFT, SURF, and DAISY achieve similar results for both datasets whereas ORB performs significantly worse.

Acc (Std.) [%]	**SIFT**	**SURF**	**DAISY**	**ORB**
ChimpZoo	**87.93 (2.28)**	86.52 (3.65)	87.58 (4.50)	53.67 (3.10)
ChimpTaï	72.46 (2.24)	**73.89 (2.88)**	69.46 (3.78)	41.11 (2.44)

Table 5.7.: ***Results of the face recognition pipeline using local features.*** The table shows the obtained accuracies and standard deviations of the proposed face recognition pipeline using different local descriptors for the ChimpZoo and the ChimpTaï dataset. The best results are printed in boldface.

As can be seen, face recognition using SIFT, SURF, and DAISY performs almost equally well on the ChimpZoo and the ChimpTaï dataset, respectively. Only ORB performs significantly worse for both datasets. The best accuracy for the ChimpZoo dataset was achieved by SIFT while for the ChimpTaï dataset SURF performed best. DAISY on the other hand achieves slightly worse mean accuracies.

However, the standard deviation across the folds for DAISY is relatively high, suggesting that the selection of training data is more crucial for DAISY than it is for SIFT or SURF. In consideration of the application presented in this thesis where fast and accurate performance might be a crucial factor, SURF is preferred as local keypoint descriptor. As claimed by Bay *et al.* in [52], SURF can be computed much more efficiently than SIFT resulting at a lower dimensional and more compact descriptor while maintaining the discriminative power of SIFT.

5.4.3. Identification Using Global and Local Features

5.4.3.1. Closed Set Identification

Previously, the proposed global and local face recognition pipelines were evaluated separately. In this section, however, it is investigated how the PRF performs if the results of both recognition strategies are combined using the decision fusion scheme proposed in Section 4.2.3.3. In particular, the robustness of the system against pose variation is examined. As discussed in Section 4.2.3, the developed identification framework is designed to recognize full-frontal faces. However, in real-world settings the system will frequently be confronted with non-frontal face data. It is thus mandatory to increase the system's robustness against off-frontal face poses as much as possible. Starting from the assumption that different features tend to misclassify different patterns, combining the results obtained by global and local features might enhance the performance of the system.

In order to thoroughly evaluate the system's performance for different poses, the off-frontal datasets of the ChimpZoo and the ChimpTaï datasets described in Section 5.2.3 were used. Example images of the three subsets *SemiLeft*, *Front*, and *SemiRight* are illustrated in Figure 5.6 while details of the datasets were given in Table 5.4. Again, a 5-fold cross-validation is used to obtain valid results. However, when testing the system's robustness against off-frontal poses, it was ensured that the system was only trained with full-frontal faces and tested on *SemiLeft* or *SemiRight* face images, respectively. Other than that all parameter settings and pre-processing steps were kept as discussed in the previous experiments.

Besides evaluating the pipeline of global and local features separately, the proposed decision-level fusion scheme is benchmarked against the fusion algorithm proposed by Gokberk *et al.* in [366]. Figure 5.18 shows the obtained mean accuracies. The according standard deviation across the folds are given in Table 5.8.

Acc. (Std.)[%]	**ChimpZoo**			**ChimpTaï**		
	SemiLeft	**Front**	**SemiRight**	**SemiLeft**	**Front**	**SemiRight**
PRF (Global)	65.86 (3.33)	92.64 (2.26)	75.64 (5.62)	53.28 (2.75)	81.68 (2.54)	51.66 (4.28)
PRF (Local)	48.51 (3.97)	86.52 (3.65)	54.85 (4.14)	38.01 (3.27)	73.89 (2.88)	36.65 (3.73)
PRF (Fusion)	**67.86 (3.88)**	**94.22 (2.25)**	**78.00 (3.58)**	**56.53 (2.95)**	**84.42 (1.41)**	**55.52 (4.19)**
Fusion [366]	58.16 (6.58)	93.51 (3.70)	71.70 (4.39)	46.23 (2.08)	79.27 (3.54)	45.20 (4.57)

Table 5.8.: ***Results of the PRF for different pose subsets.*** The table shows the accuracies and standard deviations obtained by the proposed global (PRF (Global)) and local (PRF (Local)) face recognition scheme as well as the results obtained by the proposed decision-level fusion paradigm (PRF (Fusion)). Moreover, the results are compared with the proposed decision fusion algorithm proposed by Gokberk *et al.* in [366]. The best performances were obtained by the proposed decision fusion scheme and are printed in boldface.

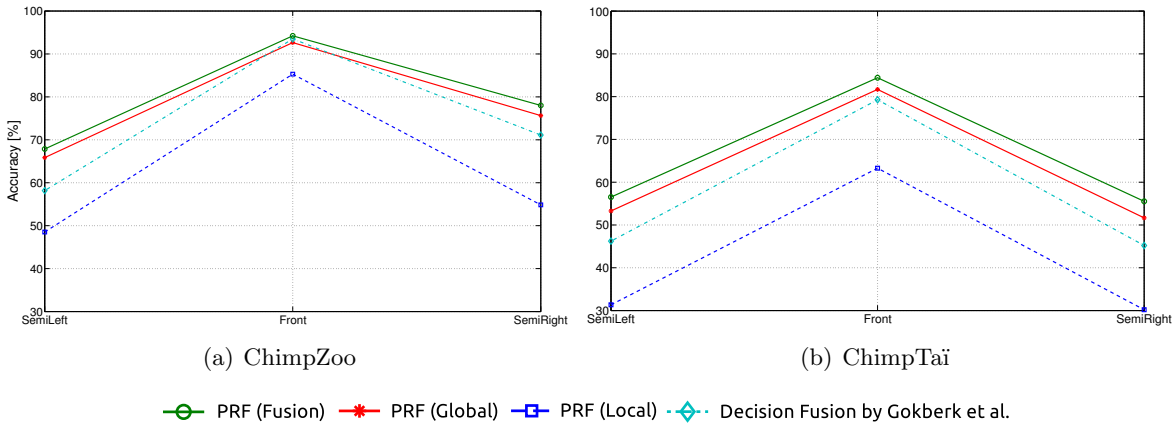


Figure 5.18.: ***Results of the PRF for three pose subsets.*** The figure shows the accuracies of the proposed PRF (green solid line) for different pose subsets of the (a) ChimpZoo dataset and (b) ChimpTaï dataset. The proposed decision fusion scheme is compared with the accuracies obtained using global (red solid line) and local (blue dashed line) features alone. Additionally, the results of the decision fusion paradigm proposed in [366] (cyan dash dotted line) is illustrated. A legend is plotted below the diagrams. Although the approach using local features alone performs significantly worse than the face recognition pipeline using global features, incorporating both modalities using the proposed decision-level fusion method enhances the system's performance for both datasets. On the other hand, the fusion paradigm by [366] which does not exhibit the classifier's confidences is not able to increase the system's robustness and thus performs worse than the proposed global face recognition pipeline.

As can be seen in Figure 5.18 and Table 5.8, the approach using local features alone performs significantly worse than the global face recognition algorithm. However, the proposed decision-level fusion scheme has a positive effect on the system's performance for frontal as well as off-frontal poses. Note, however, that the fusion algorithm proposed by Gokberk *et al.* in [366] which does not included the confidences of both classifiers is not capable to exploit the different misclassifications of both recognition strategies. Although a slight improvement for the full-frontal pose subset of the ChimpZoo dataset could be achieved, the system's performance is mostly lower than that of the global face recognition approach.

As expected, the accuracy of the face recognition system decreases notably when trained with full-frontal faces but tested with semi-profile faces. Although applying the proposed decision fusion scheme performs better than global and local features alone, a satisfying pose invariance could not be achieved. To recall from Section 4.2.3.3, the basic assumption for fusing the results of global and local features was that different features tend to misclassify different patterns. However, the obtained results indicate that almost all images that were falsely classified by the global feature approach were also misclassified by the local feature pipeline. Only a small fraction of faces that were assigned wrong by the proposed global face recognition approach were correctly classified by the alternate local identification scheme.

Thus, the improvements of fusing both recognition paradigms were only marginal compared to identification based on global features alone.

However, the results of the experiments conducted in Section 5.4.1 suggest that augmenting training data with off-frontal face data can significantly enhance accuracy for semi-profile face images while the performance for full-frontal face images is maintained. Therefore, off-frontal faces are included in the training set for the subsequent experiment in order to improve the system's robustness against pose variation. Opposed to the experiments conducted in Section 5.4.1, only real images are used for training and testing. Again, the proposed decision-level fusion scheme is compared with the global and local face recognition pipeline of PRF as well as the fusion paradigm proposed by Gokberk *et al.* in [366]. Table 5.9 lists the results obtained for the three different pose subsets when additional off-frontal data is included in the training set.

Acc. (Std.)[%]	**ChimpZoo**			**ChimpTaï**		
	SemiLeft	**Front**	**SemiRight**	**SemiLeft**	**Front**	**SemiRight**
PRF (Global)	88.73 (1.83)	91.93 (2.37)	89.47 (2.42)	76.73 (2.14)	80.98 (2.39)	75.64 (2.17)
PRF (Local)	81.09 (3.07)	85.23 (3.31)	84.05 (3.54)	69.92 (2.54)	73.88 (1.74)	71.45 (2.77)
PRF (Fusion)	**90.96 (2.33)**	**94.39 (2.23)**	**92.52 (2.81)**	**78.19 (1.85)**	**83.22 (3.77)**	**80.43 (3.25)**
Fusion [366]	87.50 (1.74)	93.14 (3.26)	87.54 (4.49)	74.67 (4.22)	79.07 (2.84)	73.48 (4.50)

Table 5.9.: ***Results of the PRF for different pose subsets by augmenting training data with additional off-frontal images.*** The table shows the accuracies and standard deviations obtained by the proposed global (PRF (Global)) and local (PRF (Local)) face recognition scheme as well as the results obtained by the proposed decision fusion paradigm (PRF (Fusion)). Additionally, the results obtained by the decision fusion scheme proposed by Gokberk *et al.* in [366] are shown. Opposed to the results shown in Table 5.8, additional off-frontal posed face images were included in the training set. Again, the best performances were obtained by the proposed decision fusion scheme and are printed in boldface.

As expected, the obtained results suggest that the system's robustness against pose variations can be enhanced significantly if both full-frontal and off-frontal face data is used for training, while the accuracy for full-frontal face data is hardly effected. Again, the proposed decision fusion scheme outperforms the other approaches for all pose subsets of the ChimpZoo and ChimpTaï dataset. As can be seen in Table 5.9, the accuracy of the proposed decision fusion scheme for semi-profile faces can be increased by up to 23% for the ChimpZoo dataset and even 25% for the ChimpTaï dataset while the high performance for full-frontal faces is maintained. Hence, the system's generalization capability can be improved significantly by including additional training data that sufficiently represents the variation to be expected in the test case.

5.4.3.2. Open Set Identification

For all previous experiments a closed-set identification scheme was applied, i.e. the number of individuals to be identified by the system is limited. In this section it is investigated how the proposed PRF performs in an open-set scenario. In an open-set identification scheme the system first has to determine if a detected face belongs to an individual in the training database and the identity of the subject only has to be reported if the subject is known. The framework is additionally tested for the open-set case because of two reasons:

1. Apart from the experiments conducted in the previous sections, where only the proposed identification scheme was evaluated, the system is now tested as a complete detection and identification framework. Thus, the detection stage will produce false positive detections which should then be rejected by the subsequent identification module.

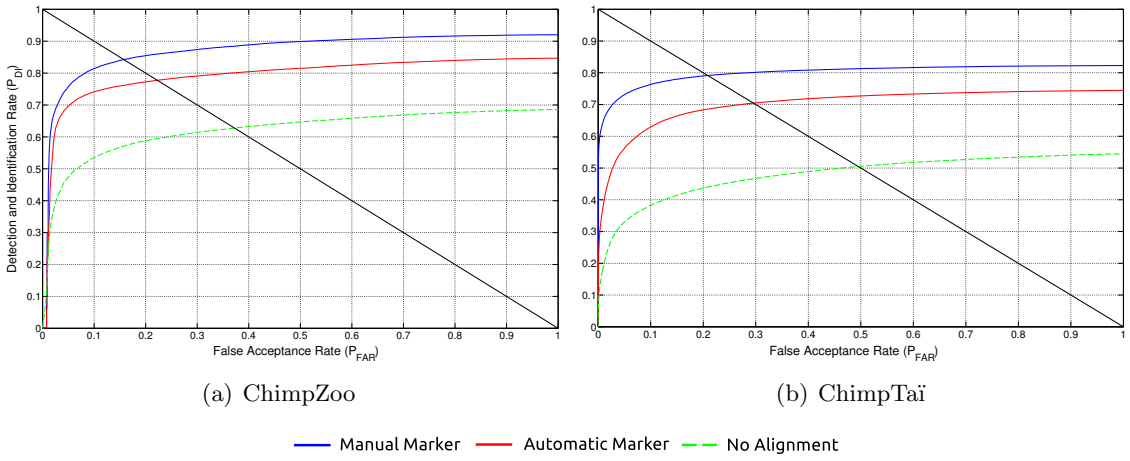
2. Often not all individuals are known a priory when applied in natural habitats for wildlife ecological research and population monitoring. Hence, individuals that have never been observed before have to be reliably rejected by the system and later be included into the training set after manual annotation.

The experimental design for open-set face recognition as well as the applied evaluation measures were reviewed in Section 5.3.2.

As already stated, the proposed PRF is evaluated as a unified framework for primate recognition including face and facial feature detection as well as identification. The algorithm described in Section 4.2.1 is first applied to the entire ChimpZoo and ChimpTaï dataset for localization of primate faces and their facial features. Detected faces with a size smaller than 64×64 pixels were ignored from further processing which drastically reduces the number of false positive detections. The remaining false positives were assigned to the class *unknown*. To test the system's capability to reject actual faces from subjects that are not included in the training set, one individual is removed from the training set in a leave-one-out manner. Thus, a 5-fold cross-validation is repeated C times, where C is the number of individuals in the training set. In each iteration an individual is removed from the training database and assigned to the unknown class such that every individual takes the role of an impostor once.

The proposed decision fusion scheme is applied for face identification since it performed best in the previous experiment. All parameters were set as described above and kept constant during experimentation. Figure 5.19 compares the ROC-curves of the proposed identification algorithm for alignment using manually annotated facial feature points (blue solid line), automatically detected markings (red solid line) and if no alignment was applied (green dashed line) for the ChimpZoo dataset (a) and the ChimpTaï dataset (b).

As already discussed in Section 5.3.2, the ROC-curve is a graphical plot which illustrates the performance of a recognition system in an open-set scenario. It plots the detection and identification rate P_{DI} versus the false acceptance rate P_{FA} by iteratively changing the rejection-acceptance-threshold τ. The according EERs and AUCs are listed in Table 5.10. Recall that the smaller the EER and the higher the AUC, the better the performance of the recognition system.



(a) ChimpZoo (b) ChimpTaï

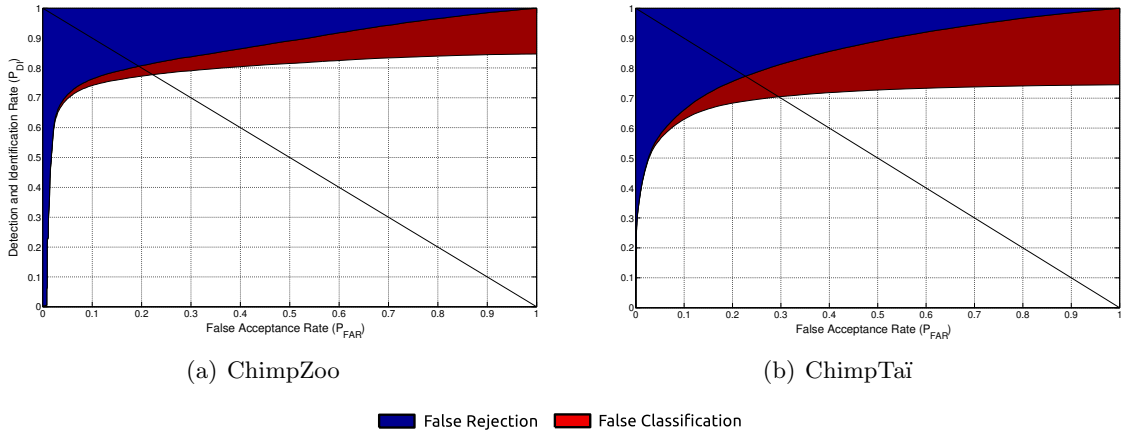
Figure 5.19.: ***ROC-curves of the PRF for alignment using different facial marker types.*** The figure shows the ROC-curves of the proposed identification system for alignment using manually annotated coordinates of both eyes and mouth (blue solid line), automatically detected facial markers (red solid line) by SHORE™, and if no alignment was applied (green dashed line). Figure (a) shows the curves for the ChimpZoo dataset, Figure (b) depicts the results for the ChimpTaï dataset. The black solid line denotes the line of equal error. As can be seen, the performance of the algorithm if manually annotated facial markings were used for alignment is better than alignment using automatically detected facial feature points. However, the accuracy of the system decreases significantly if no alignment was applied.

For alignment using manual markings the proposed algorithm performs better than for automatically detected facial fiducial points. This is because automatic localization of eye and mouth coordinates is not always as accurate as manual detection. Still, the performance of alignment using automatically detected facial keypoints significantly outperforms the proposed approach without additional alignment which supports the observations described in Section 5.4.1.

EER/AUC [%]	**Manual Markings**	**Automatic Markings**	**No Alignment**
ChimpZoo	**15.74/87.13**	22.56/79.27	37.34/62.04
ChimpTaï	**20.57/80.03**	29.58/70.13	49.54/47.93

Table 5.10.: ***Equal Error Rates (EERs) and Area under Curves (AUCs) for alignment using different facial marker types.*** The obtained EERs and AUCs for manually annotated markings are clearly better than if face alignment is performed based on automatically detected feature points. However, the performance of the algorithm again decreases significantly for the no-alignment case.

Only the relationship between correct detection and identification rate (P_{DI}) and percentage of impostors accepted by the system (P_{FA}) was investigated in the previous experiment. However, another important question is how the system's overall error rate is influenced by the other two types of errors, false rejection and misclassification. This issue is depicted in Figure 5.20 which shows the ROC-curves of the proposed system using automatically detected facial markings for ChimpZoo (a) and ChimpTaï (b). The blue area denotes the rate of false rejections (P_{FR}) while the red area shows the influence of false classification (P_{FC}) for different false acceptance rates (P_{FAR}). The lower bounds depict the ROC-curves from Figure 5.19 when automatically detected facial markers are used for alignment (red solid line). The false classification rates and false rejection rates for both datasets at the point of equal error are listed in Table 5.11.



(a) ChimpZoo (b) ChimpTaï

Figure 5.20.: ***Contribution of False Rejection Rate (P_{FR}) and False Classification Rate (P_{FC}) to the overall performance of the system.*** This figure depicts the contribution of the False Rejection Rate (P_{FR}) and the False Classification Rate (P_{FC}) to the overall performance of the proposed system when automatically detected facial features are used for alignment. Figure (a) shows the curves for the ChimpZoo dataset, Figure (b) depicts the results for the ChimpTaï dataset. The black solid line denotes the line of equal error. The blue area represents the influence of false rejections, while the red area shows the contribution of false classifications. The lower bound represents the ROC-curve obtained for automatically detected facial markers. For the ChimpZoo dataset the error rate of the system at the point of equal error is mainly caused by erroneously rejecting genuine individuals, only 3.83% is due to false classifications. For the ChimpTaï dataset, however, 18.86% of the overall error rate is caused by false rejections while 10.69% is due to misclassifications.

For the ChimpZoo dataset the main contribution to the overall error rate of the system is caused by erroneously rejected faces of genuine individuals with a P_{FR} of 18.36%. Only 3.82% of the error is due to false classifications. These results show that many facial images of known subjects were rejected as impostors because of too much pose variation, occlusion, or too challenging illumination conditions.

[%]	P_{FR} **at EER**	P_{FC} **at EER**
ChimpZoo	18.36	3.83
ChimpTaï	18.86	10.69

Table 5.11.: ***False Rejection Rate (P_{FR}) and False Classification Rate (P_{FC}) at the point of equal error.*** The error caused by misclassifications is significantly higher for the ChimpTaï dataset which is due to the higher number of individuals and the lower image quality. The influence of erroneously rejecting known individuals is almost the same for both datasets.

For the ChimpTaï dataset also the major part of the error is caused by falsely rejecting known individuals. However, the false classification rate P_{FC} is with 10.69% significantly higher than for the ChimpZoo dataset. This shows that the ChimpTaï dataset is much more challenging because it was gathered in a wildlife environment and thus contains images with a much lower visual quality compared to the ChimpZoo dataset. Furthermore, the ChimpTaï dataset contains twice as much individuals which again explains the strong influence of misclassifications to the overall error of the proposed system. Example images of incorrectly identified face images for both datasets can be found in Figure 5.21. Main difficulties arise from severe occlusion, challenging lighting conditions, and extreme facial expressions.

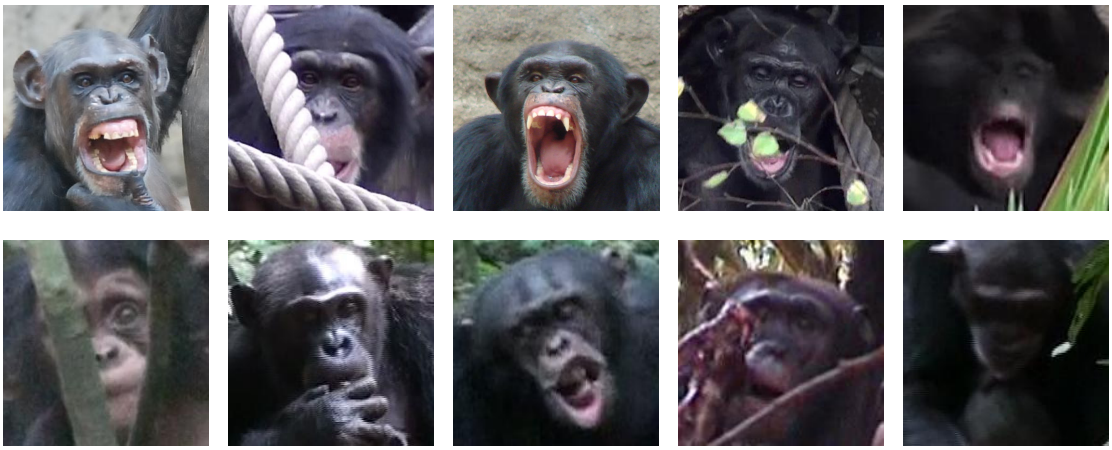
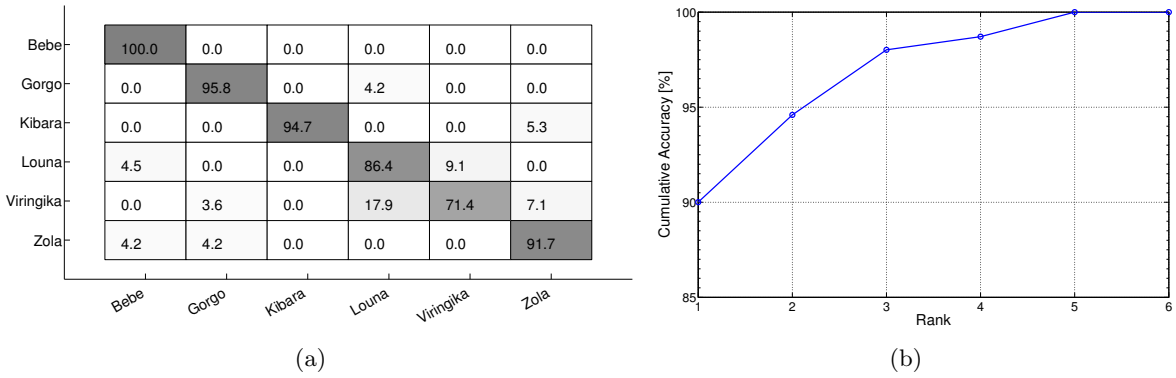


Figure 5.21.: ***Incorrect classifications of both chimpanzee datasets.*** The figure illustrates facial images that were detected by SHORE™ but could not be correctly identified the the proposed PRF. The top row shows individuals from the ChimpZoo dataset, facial images from the ChimpTaï dataset are depicted in the second row. False identifications are mainly caused by severe occlusion, difficult lighting situations and extreme facial expressions.

5.4.4. Preliminary Performance Study on Gorillas

In order to show the broad range of possible applications, the proposed identification system is also tested on a small set of captive *gorillas* gathered at the zoo of Leipzig, Germany. A detailed description of the gorilla dataset was given in Section 5.2.2.

Again, a 5-fold cross-validation was applied in order to obtain valid results. All parameters were kept as adjusted in the previous experiments and no parameter tuning was done for the experiments conducted in this section. The obtained results can be found in Figure 5.22. The confusion matrix is illustrated in Figure 5.22(a) while Figure 5.22(b) depicts the cumulative accuracy obtained by the proposed PRF.



(a) (b)

Figure 5.22.: ***Results obtained for the GorillaZoo dataset.*** The figure depicts the results of the preliminary study conducted for the GorillaZoo dataset. Figure (a) shows the confusion matrix for rank-1 accuracy while Figure (b) illustrates the graph of the cumulative accuracy.

The developed identification system for great apes is not only capable of identifying chimpanzee individuals but also other great ape species such as gorillas. The confusion matrix illustrates that most images were classified correctly. Only individual "*Viringika*" was often confused with "*Louna*" and vice versa. Some incorrect classifications are illustrated in Figure 5.23. The example images show that most misclassifications were due to severe occlusion by branches, leaves or body parts. Moreover, difficult illumination conditions and poor image quality caused by low contrast for instance are factors which might hamper the performance of the system.

By calculating the mean of the main diagonal of the confusion matrix, an accuracy of about 90% was achieved for the GorillaZoo dataset. The standard deviation across the 5 folds was 4.15%. As can be seen from the cumulative accuracy plot in Figure 5.22(b), the probability that the correct individual is among the first three proposed by the system is approximately 98%.

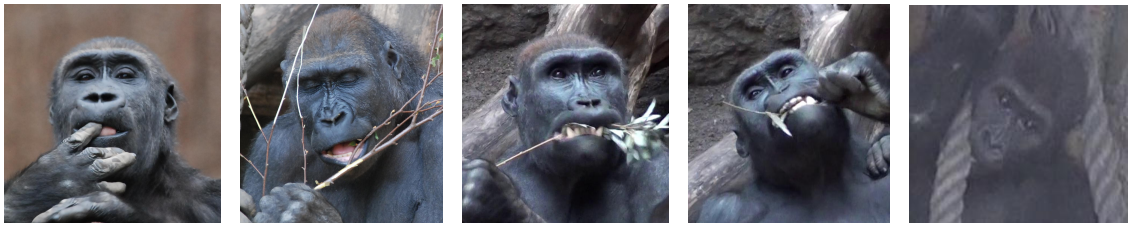


Figure 5.23.: ***Examples of incorrect classifications for the GorillaZoo dataset.*** Most of the misclassifications are caused by too severe occlusion by branches, leafs and body parts. Other reasons for incorrect recognition might be extreme facial expressions or low image quality caused by low contrast or difficult lighting.

However, it shall be noted that the GorillaZoo dataset only contains a rather small population of only 6 different captive gorilla individuals. Hence, the results discussed in this section should be considered as *preliminary* since further experiments must be conducted with larger datasets in the future in order to obtain more valid results.

5.5. Individual Identification in Videos

In the previous section the proposed PRF was solely evaluated on still images. In the past, many biological studies were based on still images or short image sequences rather than video recordings. Such acquisition protocols often had to be applied by biologists due to memory constraints necessary in wildlife settings. Nowadays, however, decreasing costs for large storage devices allow biologists to capture video sequences of endangered species with 24 frames per second (fps) or more. This trend in wildlife monitoring requires software solutions to be capable of processing this kind of data.

It was shown in Section 4.3 how face recognition algorithms for still images can be extended in order to perform identification in videos. Detected faces and facial features are first tracked through the video sequence using SHORE™ in order to assign unique IDs to every object.

The frames of the resulting face-track are subsequently sorted according to their estimated visual quality. Finally, a frame-weighting approach was proposed which weights the predictions of the F best frames in order to obtain a final classification result per face-track.
In this section, the proposed approach is evaluated using the video datasets of free-living and captive chimpanzee individuals discussed in Section 5.2.5. Models are trained on the entire ChimpZoo and ChimpTaï image datasets, respectively. Thus, no cross-validation is performed for the subsequent experiments.

Note that a thorough evaluation of each quality estimation module proposed in Section 4.3.2 is out of the scope of this thesis. Instead, the influence of selecting the best frames of a face-track prior to identification is evaluated with regard to the accuracy obtained by recognition system. For an evaluation of each quality assessment module proposed in this thesis, the interested reader is referred to [368].

Closed-Set Classification

First, a closed-set identification scheme is applied, where the system has to classify each subject as one of the training classes. Based on the observation that false-positive detections usually cannot be tracked and thus the resulting face-tracks are extremely short, the minimum length of a face-track is set to 10 frames. All tracks below that threshold are automatically classified as *unknown* and are not further processed. This procedure correctly eliminated 91.80% and 95.18% of all false-positive detections for the ChimpZoo-Video and the ChimpTaï-Video dataset, respectively.

Four different approaches are compared:

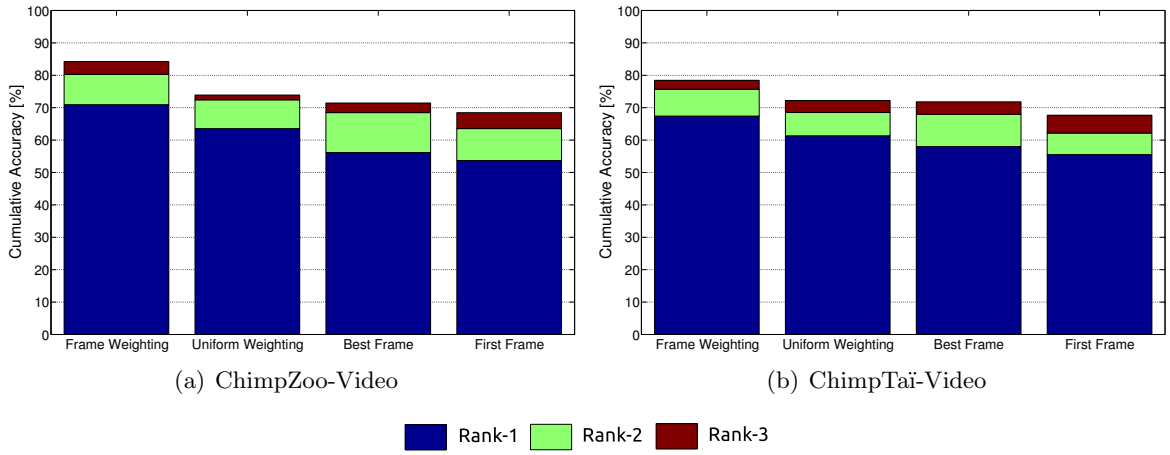
- **First Frame:** The applied face detection library SHORE™ was trained on full-frontal faces with moderate pose offsets. Hence, it can be assumed that the first frame of a face-track contains a face in full-frontal pose. Based on this assumption, the first approach identifies great apes solely in the first frame of a given face-track and remains the predicted identity for the remaining track.

- **Best Frame:** With the help of the visual quality estimation modules proposed in Section 4.3.2, the frames of a face-track can be sorted according to their suitability for face recognition. Thus, the identity of a face-track is predicted solely based on the frame with the best estimated quality in a second approach.

- **Uniform Weighting:** Based on the assumption that identification in multiple frames might increase the accuracy of the system, a uniform frame weighting scheme is applied. All frames of a face-track are first sorted according to their visual quality estimated by the proposed quality assessment modules. Then, the $F = 10$ best frames are selected for identification by applying the face recognition pipeline proposed in Section 4.2. The prediction of each frame is weighted equally, i.e. every prediction contributes the same to the final result. Thus, the uniform weighting paradigm is comparable to a maximum voting scheme where the prediction with the most votes represents the final result. In case of a tie, the prediction with the highest confidence in chosen as class affiliation.

- **Frame Weighting:** Finally, the proposed frame weighting scheme is applied. Again, the frames of each face-track are first sorted according to their estimated visual quality. Then, the $F = 10$ best frames are used for subsequent recognition. The fitted Probability Density Functions(PDFs) of the distributions of Mahalanobis-distances as well as the clusters of correct and incorrect predictions are found during training for the ChimpZoo-Video and ChimpTaï-Video dataset, respectively, as described in Section 4.3.3.

Figure 5.24 illustrates stacked bar-plots of the achieved results for all four approaches. The results including rank-2 and rank-3 recognition rates are depicted in order to get a better understanding of the functionality of each scheme. The according cumulative accuracies up to rank-3 are listed in Table 5.12.



(a) ChimpZoo-Video (b) ChimpTaï-Video

Figure 5.24.: ***Stacked bar-plots of the results obtained by PRF for videos.*** Depicted are the cumulative accuracies up to rank-3 of four different approaches for primate face recognition in video: The proposed frame weighting approach, a uniform frame weighting scheme, recognition solely based on the frame with the best quality, and identification based on the first frame of a face-track. Figure (a) shows the results obtained for the ChimpZoo-Video dataset while Figure (b) illustrates the cumulative accuracies for the ChimpTaï-Video dataset. The proposed frame weighting approach clearly outperforms the other approaches including the majority voting scheme.

Cum. Acc. [%]	**ChimpZoo-Video**				**ChimpTaï-Video**			
	Frame Weighting	**Uniform Weighting**	**Best Frame**	**First Frame**	**Frame Weighting**	**Uniform Weighting**	**Best Frame**	**First Frame**
Rank-1	**70.94**	63.55	56.16	53.60	**67.40**	61.33	58.01	55.53
Rank-2	**80.30**	72.42	68.48	63.45	**75.69**	68.51	67.95	62.10
Rank-3	**84.24**	73.90	71.44	68.38	**78.45**	72.32	71.82	67.68

Table 5.12.: ***Results of different approaches for primate face recognition in video.*** The table shows the cumulative accuracies up to rank-3 of four different approaches obtained for the ChimpZoo-Video and the ChimpTaï-Video datasets. The best results were obtained by the proposed frame-weighting approach and are printed in boldface.

For both datasets recognition solely based on the first frame performs worst. Obviously, SHORE™ is capable of accurately detecting primate faces even under difficult conditions which hamper performance of subsequent recognition. Identification is then done on a single frame basis where the quality might not be ideal for identification. Another reason for the poor performance of this approach might be inaccurate locations of facial features which are essential for face alignment. Faces, even though they are not in full-frontal pose, might be easier to detect than facial fiducial points of non-frontal faces.

The accuracy for both datasets increased by applying the proposed quality estimation modules and performing recognition on the frame with the best visual quality. This is particularly obvious for the rank-2 recognition rate which improved by 5% for both datasets. Thus, first sorting the frames according to their visual quality can increase the accuracy of facial identification quite significantly. Admittedly, the improvement of the accuracy is not as high as expected. However, when the rank-2 recognition rate is taken into account the increase in performance compared to recognition based on the first frame is quite high. This supports the assumption that recognition in a single frame often does not obtain the true individual but comes close. Thus, identification in multiple frames might lead to a better performance of the system.

As expected, the applied uniform weighting scheme performed significantly better than the previous two single-frame-based approaches. Hence, taking recognition results of multiple frames into account seems to improve the system performance significantly because identification is not dependent on a single frame with conceivably inappropriate visual quality. However, the predictions of each frame are weighted equally and the classification confidences of each frame are not taken into account. Substantial improvements were achieved by the proposed frame-weighting approach with regard to the previous three approaches. The rank-1 accuracy could be improved by more than 6% on both datasets compared to uniform weighting. The rank-3 accuracy for the ChimpZoo-Video dataset is thus 84.24% and 78.45% for the ChimpTaï-Video dataset.

Open-Set Classification

As done for face recognition in images, an open-set classification scheme was applied to test the system's capability to reject impostors while accepting and identifying genuine individuals. The results for the proposed frame weighting approach are listed in Table 5.13.

Opposed to open-set classification on still images, the performance statistics of both datasets are very similar. However, it can be noted that for the ChimpZoo-Video dataset the main contribution to the overall error comes from falsely rejecting genuine individuals while for the ChimpTaï-Video dataset the error is almost equally distributed between P_{FR} and P_{FC}.

The reason for this might be twofold: First, some tracks with insufficient quality might be falsely rejected as unknown due to the broader range of visual quality of detected faces in the ChimpZoo-Video dataset. On the other hand, face-tracks with an acceptable overall quality can be reliably identified by the proposed system. Secondly, due to the higher number of individuals in the ChimpTaï-Video dataset recognition of primates becomes more difficult. Hence, misclassifications happen more frequently which explains the relatively high false classification rate compared to the P_{FC} obtained for the ChimpZoo-Video dataset.

[%]	**EER**	**AUC**	P_{FR} **at EER**	P_{FC} **at EER**
ChimpZoo-Video	34.67	65.76	21.88	12.79
ChimpTaï-Video	35.69	63.75	18.17	17.53

Table 5.13.: ***Open-set evaluation measures of frame-weighting for both video datasets.*** This table lists the Equal Error Rate (EER) and the Area Under the Curve (AUC) of the proposed frame-weighting approach for both video datasets. Furthermore, the contributions of false rejection and false classification are listed on the right hand side of the table. While for the ChimpTaï-Video dataset P_{FR} and P_{FC} contribute almost the same to the overall error, the main part of the error for the ChimpZoo-Video dataset is caused by falsely rejecting genuine individuals.

As can be derived from the results depicted in Figure 5.24 as well as Tables 5.12 and 5.13, the developed PRF including the proposed frame-weighting approach is capable of simultaneously detecting, tracking, and identifying primate individuals in videos gathered in uncontrolled environments. However, the error caused by false classification and especially the false rejection rate is relatively high for both datasets. Furthermore, it shall be noted that the accuracies obtained by closed-set experiments for video datasets are significantly lower than those obtained for the image datasets. The main reasons for this are twofold:

1. **More challenging datasets:** The video datasets used for experimentation are more challenging than the image datasets. For instance, all videos were recorded using interlaced video. Although a deinterlacing filter was applied, remaining interlacing artifacts often hamper classification in situations were the filmed subjects move too fast. Furthermore, many face-tracks exhibit significant out-of-focus or motion blur, far off-frontal face poses, and low contrast. Such face-tracks are often hard to identify even for human beings. Representative frames of misclassified face-tracks are shown in Figure 5.25. The idea of selecting the best frame is based on the assumption that for most face-tracks there should be a high probability that an individual moves and thus generates frames which might be better suited for recognition. However, this assumption often does not hold for short face-tracks, i.e. sequences where a face could be initially detected but was lost by the object tracker after a few frames. Thus, for a significant number of face-tracks only frames

which are not well suited for recognition are part of the track. Consequently, sorting those frames based on the visual quality does not offer any advantages. Although an additional quality threshold could be introduced under which a face-track is classified as *unknown*, such a threshold was not taken into account for the experiments in this thesis. Although the accuracy would increase significantly, many true positive detections would be falsely rejected. However, in real-world scenarios, and especially for edutainment applications such a minimum quality threshold could be used to increase the system's credibility in difficult situations.

2. **More sources for error within the recognition pipeline:** The recognition pipeline for processing videos comprises three additional steps compared to face recognition in images as illustrated in Figure 4.10. A typical reason for misclassifications in still images is inaccurate alignment due to imprecise localizations of facial features. For videos, however, additionally faces and facial markers need to be tracked through the scene which might fail especially in cluttered scenes and for far off-frontal faces. Moreover, the selection of the best frames might be another source for errors. Pose estimation for instance is based on machine learning where misclassifications occur occasionally. Hence, frames with suboptimal face quality might be preferred which hampers subsequent identification. Finally, although it was shown in this section that the proposed frame-weighting algorithm outperforms single-frame-based approaches, it can be seen from the scatter-plot in Figure 4.14(a) that the clusters of correct and incorrect classifications overlap to an extend and thus cannot be clearly separated. Occasionally, even test samples with relatively good confidence measures happen to be misclassifications. On the other hand, a significant amount of correct classifications have relatively bad confidence measures due to challenging environmental conditions. Consequently, in some cases frame-weighting might assign high weights to incorrectly identified frames while others are weighted relatively low although classification was correct.

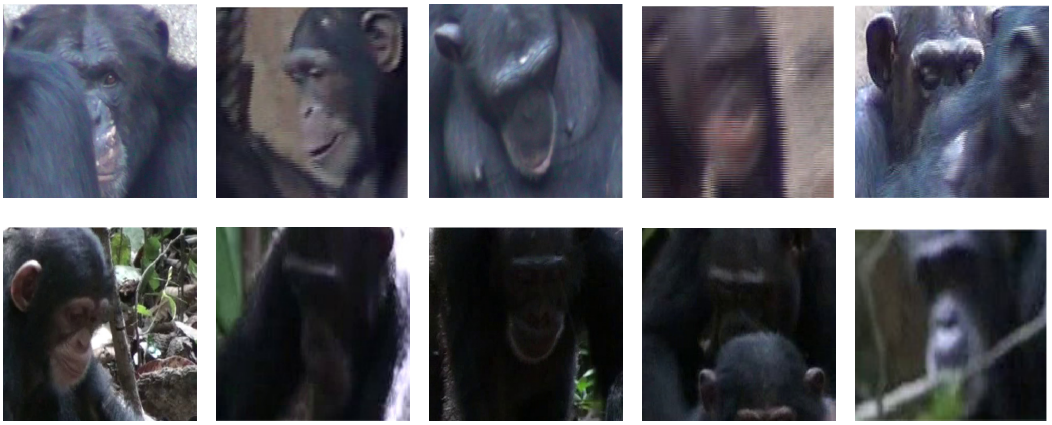


Figure 5.25.: ***Representative frames of incorrectly identified face-tracks.*** The figure shows frames of face-tracks which were misclassified by the proposed system for the ChimpZoo-Video dataset (top row) and the ChimpTaï-Video dataset (bottom row). Misclassifications frequently happen for face-tracks which were correctly detected by SHORE™ but exhibit severe occlusion, insufficient contrast, far off-frontal poses, as well as motion and out-of-focus blur. Compared to face recognition in images, the quality of face-tracks additionally suffer from video coding artifacts such as blocking or interlacing (see for instance first row, second and fourth picture). Although a deinterlacing filter was applied, faces of individuals which move too fast are often unrecognizable even for human beings.

5.6. Chapter Summary

The proposed framework for face recognition of great apes was thoroughly evaluated on image and video datasets in this chapter. Limitations and shortcomings of the proposed system were revealed and discussed; typical examples of misclassifications were illustrated and explained.

First, the annotation tool by Fraunhofer IIS was described briefly and statistics about the image and video datasets used for experimentation were given. Subsequently, the evaluation measures for closed-set as well as open-set identification used within this chapter were introduced and details about the applied experimental design were given.

During the course of the second part of this chapter, the developed Primate Recognition Framework (PRF) was evaluated on real-world image datasets which show a large variety of different poses, illumination conditions, expressions, and other extrinsic and intrinsic factors which are typical for visual footage gathered in real-world environments. The influence of different pre-processing techniques for lighting normalization and facial alignment were analyzed and the obtained performances were compared to those achieved by state-of-the-art face recognition algorithms, originally developed to identify humans. In order to show the wide applicability of the system, realistic datasets of free-living as well as captive chimpanzee and gorilla individuals were annotated by experts and used for experimentation.

Moreover, synthetic data rendered from a generic 3D model of a chimpanzee face was used to profile the investigated face recognition algorithms. Additionally, synthetic images were successfully used to augment training data in order to increase the robustness of the system against difficult illumination conditions and off-frontal poses.

The proposed extensions for robust identification of primates in video footage were evaluated within the third and final part of this chapter. The proposed frame-weighting paradigm was benchmarked against two single-frame-based approaches which identify a face-track solely based on the prediction of its first frame or the frame with the best estimated visual quality, respectively. Furthermore, a uniform frame-weighting scheme was carried out for comparison. It was shown that performing recognition on multiple frames enhances the performance of the system. It has further been demonstrated that the proposed frame weighting paradigm which weights the predictions according to the classifier's confidence significantly outperforms majority voting of multiple frames.

6. Real-World Prototype

This section gives an overview of the developed prototype for automatic face detection and recognition of primates. Although the proposed algorithms can be applied to both, images and videos, the prototype discussed in this section was mainly designed for face recognition in video since biologists mainly record short video sequences in wildlife environments. In addition to the presented Graphical User Interfaces(GUIs), a command-line interface of the proposed face recognition framework does exist which is able to process image as well as video data. The herein discussed prototype was presented at CeBIT 2012 and CeBIT 2013 and is currently used by biologists of the Max Planck Institute for Evolutionary Anthropology[1] to process video sequences of chimpanzees and gorillas gathered in national parks and zoos for behavioral studies, population monitoring, and wildlife analysis. The demonstrator is divided into three main parts which will be discussed separately in the following sections: the *Training Module*, the *Ripper Module*, and the *Graphical Interface*.

6.1. The Training Module

Figure 6.1 shows the user interface of the training module which is part of the developed Primate Recognition Framework (PRF). The training module allows the generation of training data and classification models for recognition of great apes in unseen data. The user can add new images by pressing the orange button (1 in Figure 6.1). Based on the XML annotation files which were created using the annotation tool provided by Fraunhofer IIS (see Section 5.2.1 for details), annotated faces are automatically cut out of the image. Extracted faces are presented in tabs named as the annotated individual on the right hand side of the GUI (2 in Figure 6.1). Based on the following criteria, face images are automatically selected as possible training data (faces with green frames):

- **Size:** Only face images that provide sufficient facial details in order to extract discriminating descriptors should be part of the training data. Hence, only faces with a size of at least 64×64 pixels are considered to be in the training set.
- **Vertical pose (tilt):** Faces with too large tilt angles, i.e. faces where a considerable amount of facial area is missing, should not be part of the training set. Therefore, only images annotated as *EyeLevel*, *SemiBottom* or *SemiTop* are considered.

[1] `http://www.eva.mpg.de/`. Last visit: January 29^{th}, 2014

- **Horizontal pose (yaw):** The developed face recognition system was designed for recognition of full-frontal faces. Although it has been shown that augmenting training data with off-frontal posed data has a positive effect on the system's accuracy, completely profile-posed faces should not used for training. Thus, only frontal faces with moderate yaw angles are considered as training data.

- **Occlusion:** Also occlusion by branches, leafs or other individuals are critical factors for training data generation. Therefore, only face images with *None* or *Moderate* occlusion are taken into account.

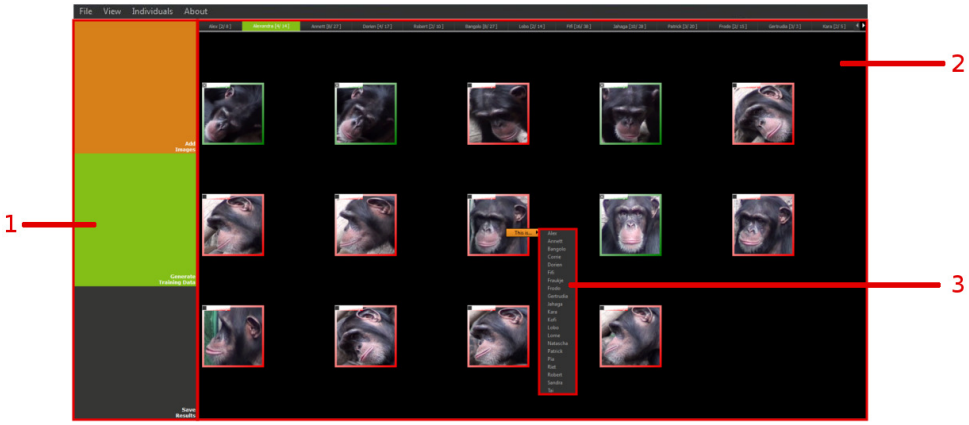


Figure 6.1.: ***Screenshot of the training module as part of the developed primate face recognition prototype.*** (1) Annotated training data can be loaded and processed in order to generate models which can later be used to recognize individuals in unseen image and video footage. (2) Based on certain criteria, a number of images are automatically selected as suitable training images (green frames) in order to train the developed algorithms. However, the operator can manually select additional faces assumed to be appropriate for training and deselect unsuitable images (red frames). (3) Face images with incorrect annotations can be reassigned to other individuals. Such images are automatically moved to the according individual tab.

Note, however, that training images are usually gathered in real-world environments. Therefore, the overall quality of training images and the number of images per individual that are necessary to span a meaningful feature space has to be traded-off against each other. Furthermore, training data which provides sufficient intra-class variation can increase the system's robustness against difficult conditions present in real-world application scenarios. Thus, the user can also manually select (faces with green frames) or deselect (faces with red frames) images. Another important option of the training module is that face images with wrong annotations can be manually assigned to other individuals (3 in Figure 6.1). After it has been decided which images should be used for training, the selected faces are aligned based on the annotated facial markings (eyes and mouth) and the proposed global and local descriptors are extracted by press-

ing the green button (1 in Figure 6.1). Furthermore, models for feature space transformation and classification are generated. Note that in order to build suitable classification models, individuals with less than 15 images are ignored and are thus not included in the training data. After feature extraction and model generation, all files necessary for applying the proposed algorithms to unseen test data are saved by pressing the gray button (1 in Figure 6.1):

1. **Training data:** A matrix of size $\mathbb{R}^{m \times l}$, where m is the size of features after projection and l is the number of training samples. This matrix is used within the SRC algorithm for identification based on global features.

2. **Labels:** A vector of size $\mathbb{Z}^l$ which contains the class labels for every training sample.

3. **Names:** A vector of strings of size C containing the names or unique IDs for every individual in the database, where C is the number of classes in the dataset.

4. **Classification models:** The SVM model used for pose estimation is fixed and does not have to be created by the training module. However, another SVM is trained based on the extracted local SURF descriptors discussed in Section 4.2.3.2. Furthermore, models for feature space transformation for local and global features are generated and saved in the training phase of Locality Preserving Projections (LPP). Moreover, the characteristic functions $\boldsymbol{\delta}_i$ (see equation 2.29 in Section 2.4.2) need be saved for every class in order to apply the SRC-based classification scheme.

5. **Mahalanobis parameters:** The sample mean and covariance matrices of the clusters of correct and incorrect classifications are used for frame-weighting. These parameters are necessary to calculate the Mahalanobis-distances of a test sample $\boldsymbol{t}$ to the cluster of correct and incorrect classifications (see Section 4.3.3 for details).

6. **Distribution parameters:** The estimated parameters of the fitted Extreme Value Distributions(EVDs) are found during training. The parameters are then applied in the test case in order to calculate the weights for the proposed-frame weighting approach.

The first four files are necessary to train the classification engine and identify known individuals or reject impostors when unseen test data is processed. Therefore, the files *training data*, *labes*, *names*, as well as the *classification models* are mandatory for the ripper module which will be discussed in the subsequent section. Files 5 and 6 are only used for the frame-weighting approach introduced in Section 4.3.3. If the operator chooses to enable frame-weighting, the *Mahalanobis parameters* and *distribution parameters* have to be specified, otherwise they are obsolete. Note that all these files are generated separately but are wrapped in an Fraunhofer IDMT specific container format named XChange Container Format (xcf).

6.2. The Ripper Module

The ripper module is the main part of the developed prototype and extracts individual data from images and videos by applying the face detection algorithm developed by Fraunhofer IIS (Section 4.2.1) and the face recognition algorithm presented in this thesis. A screenshot of the ripper module can be seen in Figure 6.2.

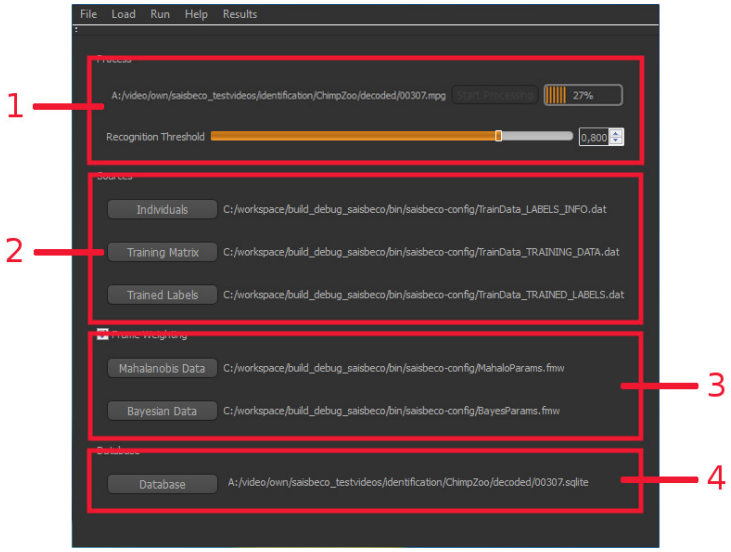


Figure 6.2.: ***Screenshot of the ripper module as part of the developed prototype.*** (1) A video file can be added by drag and drop. Furthermore the user can select the acceptance-rejection threshold, a value between 0.0 and 1.0, to reject unknown individuals. (2) Three mandatory files that contain the training data, the according labels, as well as the names of the individuals are necessary to recognize individuals in unseen data. These files can either be loaded separately of by specifying the according xcf-file. (3) If frame-weighting is enabled, two additional files have to be specified: The *Mahalanobis parameters* and the *Distribution parameters*. (4) The extracted information is saved in an SQL database. It is possible to process the same video with different parameter settings. The results are then saved into different tables within a single database.

The ripper module contains four main parts: First, the operator selects a video file to be processed by drag and drop (1 in Figure 6.2). Furthermore, the acceptance-rejection threshold can be adjusted by the user in order to reject unknown individuals not represented in the training data. The Receiver Operating Characteristic (ROC)-curves (Figure 5.20) exemplarily show how τ should be chosen to achieve a certain correct identification rate. Secondly, all necessary files (*training data*, *labels*, and the *names*) have to be loaded by drag and drop (2 in Figure 6.2). The classification models are linked with the training files in order to ensure that the correct models for feature space transformation and classification are loaded. Thus, they do not have to be specified explicitly. All files can either be selected separately or by specifying the according xcf-file container which contains all necessary files.

Additionally, the user has the option to apply the frame weighting paradigm to weight the automatically selected best frames of each face-track which was discussed in detail in Section 4.3.3. In this case, two additional files (*Mahalanobis parameters* and *distribution parameters*) have to be specified by the user. A status bar indicates how much of the video has already been processed. Once the processing is completed, the extracted metadata is written into an SQL database [406, 407] (3 in Figure 6.2) which consequently contains the following metadata:

- **Frame Number:** The number of every frame where faces were detected or tracked by SHORE™ has to be saved into the database in order to visualize the results.
- **Location:** Furthermore, the coordinates of the upper left corner of the detected region of interest as well as the width and height for each rectangle as to be saved for visualization. Moreover, the coordinates of the detected facial features, i.e. left and right eye as well as the mouth, are stored into the database.
- **Individuals:** A ranked list of the first five individuals classified by the system as well as the according confidence values are saved for every detected face. In case a face was rejected by the system, the individual is categorized as *unknown*.
- **Species:** Additionally, the species of the animal, i.e. chimpanzee or gorilla, is saved for every detected face which was classified by SHORE.

As mentioned earlier, it is possible to process the same video several times with different parameter settings in order to compare the obtained results. The results of each run can be saved independently into different tables of the same SQL database which can later be loaded by the last module, the graphical interface, to visualize the results.

6.3. The Graphical Interface

The last module is the graphical interface which visualizes the information extracted by the ripper module. The user first selects a video which has already been processed as well as the according SQL database. The graphical interface is illustrated in Figure 6.3 and comprises four main parts. The processed video is shown in the upper left corner (1 in Figure 6.3), the detected faces are marked and the best five recommendations by the proposed face recognition algorithm are displayed as sorted list above each face with confidence bars next to the names. Green bars represent high confidences while red bars represent low confidence values. Additionally, the user can start, stop or pause the video as well as load another pre-processed video. If available, additional metadata such as a mugshot, the gender, father, mother, age, etc. is displayed

in the upper right corner (2 in Figure 6.3) for every individual in the video. The lower left part of the GUI (3 in Figure 6.3) shows a timeline for every recognized individual. The white spaces represent occurrences of a certain subject. A click on a specific location in the timeline automatically causes a jump to that time slot in the video. This allows efficient browsing through videos in order to find sequences where a particular individual is present. The fourth part of the GUI (4 in Figure 6.3) shows detailed information of the current SQL database such as the name of the processed video, the number of frames, and the timestamp when the video was processed. The user is able to select a particular SQL table to visualize the results of the according run if one video was processed with different parameter settings. Moreover, the database entries can be manipulated by the user with the help of standard SQL commands to clean up the database in order to get rid of possible false detections or to correct misclassifications.

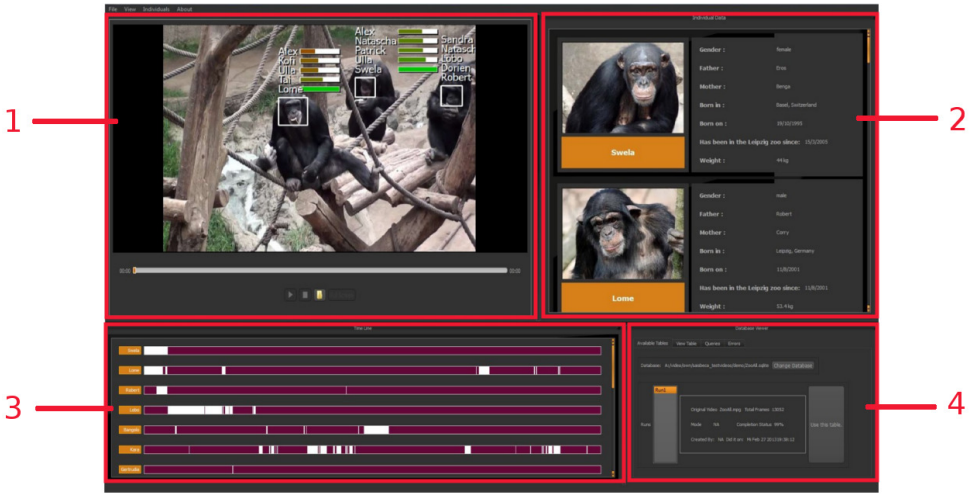
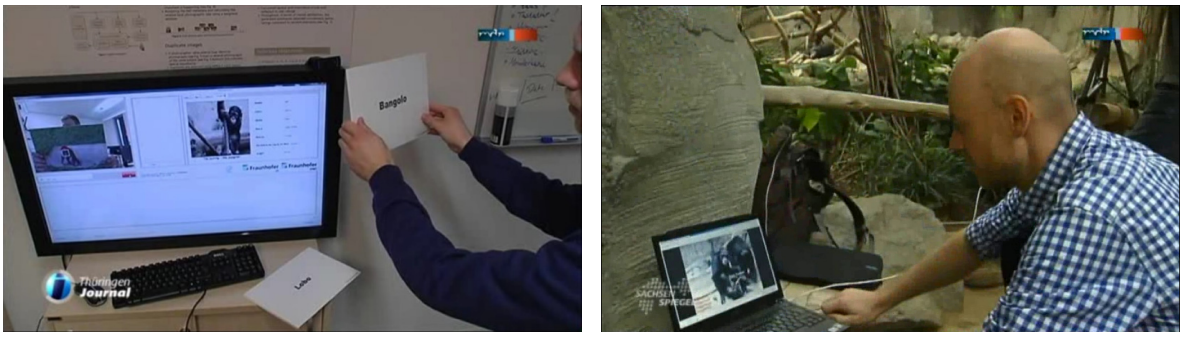


Figure 6.3.: ***Screenshot of the graphical interface of the developed PRF.*** (1) First, the operator loads a processed video and the associated SQL database. Regions of interest of localized and tracked primate faces found by SHORE™ are superimposed on the video. Furthermore, a ranked list of the first five predictions as well as the classifier's confidences are depicted for every face in terms of error bars. (2) If available, further previously gathered metainformation of individuals detected in the video can be displayed. (3) A timeline is shown for every individual in the video sequence. White bars represent time slots where a certain individual is present. The video automatically jumps to the specified time by clicking on these bars which allows quick and efficient browsing through long video sequences. (4) Information of the current SQL database is displayed. In case a video was processed with different parameter settings, the according table within the database can be selected in order to compare the results. Furthermore, the user can correct possible false detections or misclassifications by applying standard SQL commands.

Additionally, the graphical interface contains a live-mode if an external camera device is attached to the computer. This mode can be used for near real-time detection and identification of primates. For demonstration, faces of individuals living in the zoo of Leipzig were printed on signboards and held in front of the camera to show the performance of the developed technology.

Moreover, preliminary experiments were successfully conducted in *Pongoland*[2], a theme-world within the zoo of Leipzig, where promising results could be achieved. Hence, besides assisting researchers with tedious annotation work of remotely gathered videos of primates in national parks, a second application scenario of the Primate Recognition Framework (PRF) could be a near real-time capable biometric supervision system for great apes living in zoos or wildlife parks. A remote camera in combination with a control device allows visitors to monitor great apes and get an immediate feedback of the filmed individual on a screen. Such an interactive edutainment station would be a modern and novel application of a visual animal biometric system and could thus offer visitors first-hand experience of state-of-the-art technology. Moreover, the developed PRF has the potential to give the wider public deeper insights into ecological and social structures of captured great ape populations and thus contribute to the awareness of nature conservation and species protection. Furthermore, a number of biological studies regarding behavioral ecology and communicative complexity for instance are conducted with captive primates in the zoo. The presented PRF might further be used by biologists to design and conduct new experiments which might eventually help to answer a number of current biological questions and thus help researches to obtain new insights to the history of humankind. Figure 6.4 shows the demonstration of the live-mode with face images printed on signboards (a) and the preliminary real-world experiments conducted at the zoo of Leipzig, Germany (b).



(a) (b)

Figure 6.4.: ***Demonstration of the live-mode of the Primate Recognition Framework (PRF).*** (a) Recognition of a chimpanzee face printed onto a signboard. The face is automatically detected by SHORE™ and subsequently recognized using the face recognition algorithms presented in this thesis. (b) Setup of the preliminary experiments conducted in *Pongoland*, zoo of Leipzig, by Hjalmar Kühl. The proposed primate recognition framework can be used for near real-time detection and recognition of captive individuals which could eventually lead to an interactive hardware-software demonstrator for zoo visitors as well as to an automatic analysis tool for behavioral researchers. [I06, I07]

[2] `http://www.zoo-leipzig.de/themenwelten/pongoland/` Last visit: February 2^{nd}, 2014

7. Conclusion and Future Work

7.1. Thesis Summary

In the ongoing biodiversity crisis, many species including great apes such as chimpanzees or gorillas are on the brink of extinction and need to be protected. An essential part of efficient biodiversity and wildlife conversation management is population monitoring and individual identification in order to estimate population sizes, asses viability and evaluate the success of implemented protection schemes. Monitoring techniques using autonomous recording devices have been used extensively by biologists and gamekeepers for that purpose. However, manually processing large amount of data is tedious work and therefore extremely time consuming, highly cost intensive, and error prone.

To overcome these issues, this thesis proposed a unified framework for automatic detection and identification of captive and free-living great apes in image and video footage acquired in real-world environments. To the best of the author's knowledge, a fully-automatic system for individual identification of great apes has not been proposed within the research community before. Based on the assumption that humans and great apes share similar properties of the face, algorithms originating from human face recognition were adapted and extended in order to identify great apes in their natural habitats. Realistic image and video datasets of chimpanzees gathered at the zoo of Leipzig, Germany and the Taï National Park, Côte d'Ivoire, Africa were annotated by experts to provide realistic benchmark datasets. The system was thoroughly evaluated and compared against state-of-the-art human face recognition systems and it was shown that the proposed Primate Recognition Framework (PRF) outperforms the competing approaches on all datasets. The developed framework was furthermore successfully tested on a small dataset of captive gorillas to show the wide applicability of the proposed algorithms. Moreover, in order to provide a proof of concept, the proposed algorithms were implemented in a prototype which can be used by biologists, ecologists, and gamekeepers to automatically identify primate individuals in their natural habitats. Due to its near real-time capability, another novel applicability of the proposed PRF might be an interactive tool for individual identification of great apes in zoos or wildlife parks. As a result, the developed system for non-intrusive wildlife observation might not only offer visitors the possibility to get insights into ecosystems and the social complexity of chimpanzee or gorilla groups but also contribute to the public awareness of nature conservation and species protection.

In summary, the main contributions of this thesis are as follows:

- **Literature Survey**: A comprehensive literature review of existing visual animal biometric systems was given in **Chapter 3**. Advantages and shortcomings of proposed algorithms were discussed with respect to the task at hand. Since the proposed framework is based on face detection and recognition technology a broad overview of facial biometric systems was presented as well.

- **Proposed Framework**: The proposed algorithms to automatically detect and recognize primate individuals in image and video footage were explained in **Chapter 4**. The third party library SHORE™, developed and extended by Fraunhofer IIS, was applied to detect faces and facial features of great apes. A simple but effective face alignment strategy based on affine transformation was proposed in order to ensure comparability of extracted visual descriptors for all training and test instances. Starting from the assumption that different features tend to misclassify different patterns, great apes were identified by fusing the results of a global and local face recognition pipeline, both based on complementary descriptors. Additionally, the proposed system was extended to recognize individuals in video sequences. Faces were first tracked through the sequence using SHORE™. The extracted face-tracks were then analyzed according to their *visual facial quality* in order to detect the frames that are most promising for subsequent identification. Moreover, a novel frame-weighting approach was proposed to combine the identification results of multiple frames into a single final prediction per face-track.

- **Evaluation**: The proposed identification framework was thoroughly evaluated on two realistic image and video datasets of captive and free-living chimpanzee individuals gathered in real-world environments in **Chapter 5**. The influence of different pre-processing methods was evaluated and advantages as well as limitations of the proposed PRF were discussed. It has further been shown that augmenting training data with synthetic face images rendered from a generic 3D model of a chimpanzee face can increase the system's robustness against pose-offsets and different lighting conditions to a certain extent. Moreover, a preliminary study for captive gorilla individuals was conducted to demonstrate the wide applicability of the developed algorithms.

- **Real-World Prototype**: A real-world prototypical implementation was developed to aid non-intrusive biomonitoring of great apes. The system is currently used by biologists, ecologists, and gamekeepers to automatically analyze image and video footage gathered in natural habitats of great apes. A first study which utilizes the developed software was successfully conducted in [408], where the site use of two unhabituated chimpanzee groups was estimated using a manual and a semi-automatic approach. Biologists found that semi-automated data processing required only 4% of the time compared to an entirely manual

analysis in order to obtain comparable results. The authors conclude that the developed PRF shows great potential in providing assistance for annotation of camera trap data. **Chapter 6** gives an overview of the developed prototype which comprises three main components: the *Training Module* which can be used to train the identification algorithm, the *Ripper Module* which uses the previously trained model to automatically process image and video footage, and the *Graphical Interface* which visualizes the obtained results.

7.2. Limitations and Future Work

Despite the sound performance achieved by the proposed framework on challenging datasets, the acquired footage can only reflect a fraction of the true variance present in natural habitats of great apes. Although cross-validation was applied to obtain valid results, the generalization capability of the developed identification system can only be approximated to some extent. The results presented in Chapter 5 should therefore be seen as a rough estimate of what could be achieved by the proposed system when applied in real-world settings.

Figures 5.21, 5.23, and 5.25 show images that were regularly misclassified by PRF and hence provide insights into limitations of the system. The most frequently occurring categories of false classifications and ideas how the techniques proposed in this thesis could be extended in the future to possibly overcome these limitations are discussed below.

- **Pose Variation, Facial Expressions, and Illumination**: Most of the incorrect classifications are due to far off-frontal poses, extreme facial expressions, and high dynamics in lighting. A number of pre-processing methods were proposed in this thesis to overcome these difficulties to some extent. It was shown for instance in Section 5.4.1 and 5.4.3 that additional off-frontal posed training data can increase the system's robustness against pose variation. Another attempt to overcome these limitations for video sequences was to use an automatic pose estimation approach to select the frames which are best suited for subsequent identification. Still, faces with too large tilt or yaw angles are often rejected or incorrectly identified. Recent success in 3D face modeling however suggests that more advanced alignment strategies could be used to further enhance the systems robustness against pose variation [342]. Another possibility to resolve the ambiguities resulting from large pose variations is to train multiple models for identification. After automatic head pose classification the one model which was trained on facial data that best represent the actual pose could be used for identification. Furthermore, other factors for misclassification are large appearance variations caused by extreme facial expressions as well as occlusion by branches, leafs, or conspecifics. Local descriptors extracted around eyes and nose were applied in this thesis to reduce the effect of these type of distortions. Although it was

shown in Section 5.4.1 that the proposed ELGTPHS descriptor is relatively robust against moderate illumination changes, extreme lighting conditions such as dark shadows within a face are particularly challenging, especially for dark pigmented faces which usually already exhibit a low contrast. It has recently been shown that general invariance against pose, illumination, and facial expressions can be achieved by multilinear tensor-based approaches where every mode of a tensor represents different image variations [409, 410]. Multilinear interactions of different factors such as pose, lighting, and expression can thus be efficiently modeled to enhance the robustness of face recognition systems. However, a drawback of such an approach is that huge amounts of training data are required to adequately represent different extrinsic and intrinsic factors which can hardly be achieved in natural settings. A possible solution to overcome the burden of gathering adequate training images could again be 3D modeling of primate faces. The results presented in Section 5.4.1 indicate that synthetic training data rendered from a generic 3D model can be used as training data. Different lighting conditions, poses and - if more elaborate techniques for 3D face modeling could be applied - even facial expressions can be generated and included into the training set. Recommended future work therefore includes determining to what extent models trained on artificial data represent the variation present in real datasets. In the long run, the final goal should be to train a face recognition system with a minimum amount of real data. Thus, it is intended to further explore the application of 3D modeling techniques to one-shot learning: Using only one real image of an individual to seed synthesis for recognition in a range of image situations.

- **Environmental Clutter**: Cluttered background is particularly challenging for the face detection stage of the proposed PRF. Especially for face detection in still images false positive detections are often caused by cluttered background present in natural habitats of great apes. Bootstrapping was used within the training stage of SHORE™ to increase the system's robustness against cluttered background, where false positive detections found on a validation set were included into the training set as negative examples. As shown in [343], the number of false positive detections for face detection in videos can be significantly reduced by the proposed tracking-by-detection approach. It was shown in Sections 5.4 and 5.5 that most of the remaining false positive detections can be rejected by the subsequent identification module using an open-set identification scheme. However, more advanced tracking techniques such as *particle filters* [163] might further reduce the false positive rate and should therefore be investigated in the future. On the other hand, Sandwell and Burghardt showed in [126] that Deformable Part-Based Models (DPMs) can be used to reliably detect off-frontal posed chimpanzee faces while being less sensitive to environmental clutter.

- **Video Recording Errors**: Another source for incorrect classifications are errors resulting from the hardware used for autonomously record videos in tropical forests. Typical video recording artifacts such as *interlacing*, *blocking*, and *motion blur* might hamper individual recognition of animals. Software modules for quality assessment were proposed in Section 4.3.2.2 to automatically select frames within an extracted face-track before recognition. Furthermore, a proper set-up of modern recording devices in the field and a well-developed field acquisition protocol are necessary to overcome these difficulties in the future. Probably the best way to develop such a protocol is to evaluate different devices, locations, and recording conditions to simultaneously optimize animal capture rates and visual data quality. First attempts to optimize data acquisition in natural habitats have already been done in [33] but are continuously improved by biologists.

A selection of additional directions of research and practical implementations that could be pursued in order to extend the functionality of the proposed framework in the future are briefly discussed below.

- **Unsupervised Face Clustering**: One of the biggest challenges in building models for individual identification is the difficulty to immediately obtain adequate annotations. Acquisition of training data and ground-truth annotation still has to be performed by human experts which is not only time consuming but also tedious routine work. Unsupervised clustering techniques could be applied as a preliminary processing stage to overcome these issues where derived feature vectors of multiple individuals are grouped in an unsupervised fashion according to their similarity. The challenge here is that the number of individuals is not known in advance. Therefore, traditional clustering methods such as *k-means* for instance are not suitable for that purpose. However, more elaborate techniques such as *g-means* [411], *Affinity Propagation* [397], or *Rank-Order Distance Based Clustering* [398] could be used for that purpose since remarkable results could be achieved for human face datasets in the recent past. Preliminary results of these automatic clustering techniques applied to the facial databases of primates are promising [1, 412] but more intensive research needs to be done in the future. Presumably, semi-automatic approaches are most promising for that task. Once the individuals are automatically clustered into several exclusive groups, user feedback could be used to optimize the clustering results.

- **Hierarchical Approaches**: In many situations it might be beneficial to first solve easier sub-problems rather than identify the occurring animal directly. For many applications these sub-problems might have a hierarchical structure. For instance in case of primate recognition the detected apes could first be clustered according to the species, e.g. chimpanzee, gorilla, orang-utan, etc. Within each species a second layer could categorize the

detected animals according to their gender. Additionally, different age groups could be taken into account to limit the number of individuals to be identified. Such hierarchical interpretation allows users to provide their feedback on different hierarchical levels. In the ideal case, the optimization via user feedback leads to the optimal clusters containing single individuals. When the clusters are optimized and cannot be improved by user feedback, data belonging to one cluster could finally be used to train individual models.

- **Super Resolution**: It was shown in Section 5.5 that the proposed frame-weighting scheme is capable of significantly enhancing the performance of the system compared to approaches that perform recognition on a single frame. However, quality and resolution of video recordings acquired in natural settings might sometimes not be sufficient to perform robust and accurate recognition. Moreover, in typical surveillance scenarios, cameras are often at a considerable distance from the subjects and the captured image typically contains only a small region enclosing the subject's face. So called multiframe super-resolution methods [413, 414], a class of techniques developed to enhance the resolution of imaging systems, might be utilized in future applications in order to improve the system's performance for low-resolution video recordings. Although super-resolution has extensively been studied in the past, applications on real-world video still remains challenging due to oversimplified motion models or other assumptions that do not hold in reality [415]. However, recently developed super-resolution techniques designed with a special focus on face recognition applications [416, 417, 418], also known as *Hallucinating Faces* [419, 420], appear to be promising and might therefore lead to a more robust and accurate primate recognition framework for low-resolution footage in the future.

- **Advanced Feature Sets**: It has been shown that the combination of ELGTPHS and SURF provides descriptive information for identification of primates in real-world environments. However, a variety of other feature types were proposed in the recent past which could be used to extract additional information which might be helpful for robust and accurate identification. A number of sophisticated local descriptors have been proposed in the recent past which have been shown to perform better than standard SIFT and SURF under certain conditions [404, 230, 233]. In addition to local keypoint descriptors, novel texture descriptors were proposed recently which are based on fractal geometry [421].

  Nowadays, multifractal analysis is considered to be a promising tool for image processing, in particular for texture recognition [422, 423, 424]. In the context of visual animal biometrics, those feature sets could serve as additional tools for description and comparison of complex individual patterns and should therefore be investigated more thoroughly in the future.

In broader terms, an integration of the probabilistic output of the developed PRF in existing monitoring frameworks is necessary in the future to proof the feasibility of animal biometric systems for automatic population monitoring and ecological studies. Therefore, statistical approaches need to be developed to derive visitation rate estimates, occupancy rates, density and abundance estimates based on camera trap data and the developed primate identification system. Another venue to extend the approaches proposed in this thesis is to automatically quantify communication behavior and their physiological correlates in human and non-human primate social systems in natural settings by means of automated multi-modal tools. Although it has been shown in this thesis that algorithms for automatic facial detection and recognition are promising for non-invasive monitoring of great apes, they are limited to image and video footage. Different sensor types (e.g. audio, video, infrared imaging, etc.) should be used in the future to gather multiple types of information to adequately address the multi-modal aspect of animal biometrics. Consequently, algorithms should be developed in the future that process multiple sensory inputs in a truly multi-modal fashion. Especially for automatic primate communication analysis the combination of audio and video analysis techniques might lead to a more flexible and therefore more robust system.

Although promising results could be obtained by the proposed PRF on real-world datasets of captive and free-living chimpanzee and gorilla individuals, a completely automatic approach will not be capable of entirely replacing humans in the near future. However, simple annotation tasks do not necessarily need to be performed by professional scientists. Therefore, *citizen science* has become a mainstay in biodiversity research and has thus increased the scale of ecological field studies tremendously [425]. According to the Oxford English Dictionary, citizen science, also known as *crowd-sourced science*, is defined as *"[...] scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions [...]"* [426]. Biologists from the Max Planck Institute for Evolutionary Anthropology recently launched a citizen science project called *"Chimp & See"*[1] in collaboration with Zooniverse, one of the largest, most popular, and most successful web portals for citizen science. In this project, volunteers watch videos taken by various camera traps in Africa, identify chimpanzees, and annotate their behavior in order to help researches to learn more about human evolution. Such an approach in combination with the proposed algorithms for automatic detection and identification of great apes has the potential to contribute significantly to wildlife research and ecological science in general.

[1] `http://www.chimpandsee.org/beta/#/classify` Last visit: April 7^{th}, 2015

7.3. Concluding Remarks

Biologists and gamekeepers recently started to use remote cameras and audio recording devices for wildlife monitoring in order to overcome the catastrophic decline of biodiversity [427, 428, 33]. The objective of the emerging field of *Visual Animal Biometrics* is to develop and apply approaches to automatically detect, represent, and interpret phenotypic appearance of various animal species to overcome the burden of manually annotating image and video footage [85]. While many existing approaches focus on patterned animals, it has been demonstrated in this thesis that African great apes such as chimpanzees and gorillas can be reliably detected and identified by their facial appearance. Face recognition algorithms, originally developed for human identification, were adapted and extended for the challenging task of automatically identifying primates in their natural habitats.

The proposed system is capable of reliably detecting and identifying primate individuals although evaluation data was acquired in real-world environments which place high demands on the system due to a broad variety of illumination, partial occlusion, non-cooperative subjects, and off-frontal head poses. Furthermore, it is possible to reject subjects which are not present in the training database by adjusting an acceptance-rejection threshold accordingly. Hence, the proposed framework can be applied for identification of great apes in real-life scenarios and thus might provide substantial assistance for tedious annotation work of gathered images and videos.

Therefore, the author of this thesis believes that the developed system has the potential to open up new venues for efficient and innovative wildlife monitoring and biodiversity conservation management. Intensive pilot studies are currently conducted in Loango National Park, Gabon, [33] and Taï National Park, Côte d'Ivoire, [428] using autonomous infrared-triggered remote video cameras in combination with the presented biometric identification software. These studies already provided promising results, demonstrating the potential of such an approach for biomonitoring. For instance, a study by Crunchant *et al.* [408] successfully applied the developed PRF to estimate the site use of two unhabituated chimpanzee groups. The semi-automatic approach which took advantage of the proposed system was compared with a more traditional method which was based on manual annotation of camera trap data. The biologists found that semi-automated data processing saves a significant amount of time and human resources without compromising the quality of the results. Hence, the study by Crunchant *et al.* demonstrates the effectiveness of the proposed PRF. The authors encourage other researchers to utilize automated visual animal biometric software to increase the usefulness of collected data from field studies.

Moreover, a real-time capable edutainment application to identify captive chimpanzees in zoos or wildlife parks is currently under development. Such an interactive tool offers zoo visitors the possibility to monitor great apes with a remote camera device and get an immediate feedback about individual information as well as external metadata about the filmed subject. The author of this thesis believes that such an interactive edutainment station could offer first-hand experience of state-of-the-art technology presented in a novel application. More importantly, with the help of the proposed PRF, zoo visitors might get insights into ecological and social structures of primate populations quite intuitively. Moreover, designed as a *serious game*, visitors could collect data which might be useful for biologists for behavioral ecological research or to analyze communicative complexity of captive chimpanzee groups. Hence, apart from a professional tool for biologists and gamekeepers the proposed framework could also be applied by non-experts in a citizen-science fashion and thus help researches to obtain new insights into social systems of primates and ultimately investigate the history of humankind.

Appendix A.

Symbols and Notation

x	scalar or continuous variable	$\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$	column vector	$\boldsymbol{X}$	matrix
$\mathcal{X}$	graph or kernel function	$\{x_1, x_2, \ldots, x_n\}$	a set	$\boldsymbol{X}^{\mathrm{T}}$	transposed of $\boldsymbol{X}$
$\boldsymbol{X}^{+}$	Moore-Penrose pseudo-inverse of $\boldsymbol{X}$	$\boldsymbol{X}^{-1}$	inverse of $\boldsymbol{X}$	$\boldsymbol{X} \cdot \boldsymbol{Y}$	dot product
$\boldsymbol{X} \odot \boldsymbol{Y}$	Hadamardt-Schur product	d	differential operator	∂	partial operator
$\mathbb{R}$	real numbers	$\mathbb{N}$	natural numbers	$\mathbb{B}$	binary numbers
x^y	exponentiation	$\log_b x$	logarithm with basis b	exp	exponent operator
$\sqrt{x}$	square-root of x	$\|\boldsymbol{x}\|_p$	ℓ^p-norm of $\boldsymbol{x}$	$\|x\|$	absolute value
π	Pi, Ludolphian number	e	Euler's number	λ	Eigenvalue
i	imaginary unit	$\Re\{x\}$	real-part of x	$\Im\{x\}$	imaginary-part of x
arg	argument operator	$\mapsto$	mapping operator	Δ	delta operator
$*$	convolution operator	∞	infinity	$\times$	product operator
$=$	equality	$<, >, \ll, \gg$	strict order signs	$\leq, \geq$	order symbols
$P(A\|B)$	conditional probability of A given B	$P(A)$	probability of event A	$\mathcal{F}\{\cdot\}$	Fourier-transform
$\sum$	sum operator	$\prod$	product operator	min	minimum element
max	maximum element	$\sin(x)$	sine of x	$\cos(x)$	cosine of x
$\forall$	for all	$x \in X$	x is in X	$\sup\{\boldsymbol{x}\}$	supremum of $\boldsymbol{x}$

Appendix B.

Parameters for Face Recognition Using Global Features

	Parameter	Description	Value
Feature Extraction			
Gabor Wavelets	H	Width/height of the Gabor kernels	31
	p	Energy preserving ratio of the Gabor kernels (eq. 4.5)	0.9
	k_{max}	Maximum frequency of the Gabor kernels (eq. 2.1)	$\frac{\pi}{2}$
	f	Spacing between kernels in the frequency domain (eq. 2.1)	$\sqrt{2}$
	σ	Ratio of the Gaussian window to the wavelength (eq. 2.1)	π
	F	Number of rotations	8
	S	Number of scales	5
Extended Local Ternary Patterns (ELTP)	B	Number of blocks each Gabor Magnitude Picture (GMP) is divided into	3×3
	P	Number of neighboring pixels (eq. 2.3)	8
	R	Radius around the center pixels (eq. 2.3)	2
	α	Parameter that scales the standard deviation of the image patch for thresholding (eq. 4.7)	0.2
Feature Space Transformation			
Locality Preserving Projections (LPP)	m	Size of the feature space after projection (eq. 4.8)	160
	σ	Scaling parameter of the heat-kernel (eq. 4.9)	100
Classification			
Sparse Representation Classification (SRC)	τ	Real-valued non-negative parameter (eq 4.13)	$0.1\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}\|_\infty$

Table B.1.: ***The default parameters of the proposed face recognition pipeline using global features.*** All these parameters remained constant for the experiments conducted in this thesis.

Appendix C.

Parameters for Face Recognition Using Local Features

	Parameter	Description	Value
Feature Extraction			
U-Speeded-Up Robust Features (SURF)	–	Ratio of the width of the region of interest to the width of aligned image	1:7
Feature Space Transformation			
Locality Preserving Projections (LPP)	m	Size of the feature space after projection (eq. 4.8)	160
	σ	Scaling parameter of the heat-kernel (Eq. 4.9)	100
Classification			
Support Vector Machine (SVM)	C	Penalty parameter (Eq. 2.31):	
		Start value for grid search	2^{-5}
		Final value for grid search	2^{15}
	γ	Parameter for creating an Radial Basis Function (RBF) kernel (Eq. 4.14):	
		Start value for grid search	2^{-15}
		Final value for grid search	2^{-3}
	Step sizes	coarse grid	2
		middle grid	0.5
		fine grid	0.25

Table C.1.: ***The default parameters of the proposed face recognition pipeline using local features.*** All these parameters remain constant for the experiments conducted in this thesis. Note that the parameter configuration of C and γ is found by applying a 3-step coarse-to-fine grid search.

Appendix D.

Results for Different Face Alignment Strategies

Acc (Std.) [%]	NONE	ROTATE	AFFINE
PRF (global)	**75.69 (3.10)**	**91.60 (3.43)**	**92.65 (2.26)**
Eigenfaces	36.53 (4.19)	44.58 (1.96)	51.04 (3.65)
Fisherfaces	25.17 (2.53)	60.13 (5.94)	74.47 (3.66)
Laplacianfaces	44.05 (2.68)	69.40 (4.70)	80.76 (5.69)
Randomfaces	34.96 (5.54)	52.97 (2.14)	60.83 (4.07)
BDCT	39.33 (2.71)	68.18 (1.84)	78.84 (1.67)
GSRC	60.13 (1.47)	72.72 (4.69)	82.34 (3.49)
RSC	43.00 (6.18)	65.03 (1.53)	72.20 (3.93)
RRC	35.83 (4.70)	61.01 (2.92)	66.25 (2.99)

Table D.1.: ***Results of state-of-the-art algorithms combined with three alignment strategies for the ChimpZoo dataset.*** The table shows the obtained accuracies and standard deviations of the experiments described in Section 5.4.1 for the ChimpZoo dataset. The best results are printed in boldface and were obtained by the proposed global face recognition pipeline as part of the Primate Recognition Framework (PRF).

Acc (Std.) [%]	NONE	ROTATE	AFFINE
PRF (global)	**65.28 (3.99)**	**78.44 (2.08)**	**81.68 (2.54)**
Eigenfaces	50.40 (3.20)	43.01 (4.14)	47.57 (3.42)
Fisherfaces	20.14 (1.80)	11.84 (1.89)	6.07 (1.72)
Laplacianfaces	45.95 (4.95)	50.70 (3.72)	56.27 (3.28)
Randomfaces	42.61 (4.17)	47.67 (3.41)	52.53 (1.02)
BDCT	44.73 (3.53)	57.69 (1.58)	58.50 (2.36)
GSRC	59.00 (2.63)	62.14 (2.80)	68.82 (2.57)
RSC	49.69 (4.33)	55.26 (3.79)	58.70 (2.27)
RRC	37.75 (5.47)	50.70 (4.18)	54.04 (2.95)

Table D.2.: ***Results of state-of-the-art algorithms combined with three alignment strategies for the ChimpTaï dataset.*** The table shows the obtained accuracies and standard deviations of the experiments described in section 5.4.1 for the ChimpTaï dataset. Again the best results are printed in boldface and were achieved by the proposed global face recognition pipeline as part of the PRF.

Appendix E.

Results for Different Illumination Normalization Algorithms

Acc (Std.) [%]	NONE	HIST	CLAHE	MSR	IMADJUST	GAMMA + DOG
PRF (global)	**92.66 (2.26)**	**91.61 (1.58)**	**91.26 (2.83)**	**91.08 (2.76)**	**93.18 (1.74)**	**92.48 (4.21)**
Eigenfaces	51.05 (3.65)	63.11 (8.73)	69.76 (6.33)	55.07 (1.96)	61.01 (4.36)	56.99 (3.32)
Fisherfaces	74.48 (3.66)	84.09 (3.10)	83.22 (4.08)	82.87 (4.47)	80.24 (2.94)	76.75 (3.83)
Laplacianfaces	80.77 (5.69)	85.84 (3.48)	86.54 (3.54)	83.39 (3.25)	83.22 (7.34)	80.42 (3.80)
Randomfaces	60.84 (4.07)	73.60 (4.18)	65.21 (5.14)	60.66 (3.42)	66.78 (4.75)	55.42 (4.13)
BDCT	78.85 (1.67)	76.57 (5.28)	78.67 (2.58)	76.92 (3.80)	78.67 (4.44)	76.57 (3.39)
GSRC	82.34 (3.49)	85.66 (3.41)	87.76 (4.07)	81.82 (3.89)	82.34 (4.03)	81.64 (3.04)
RSC	72.20 (3.93)	83.39 (4.17)	78.67 (3.34)	72.55 (3.91)	74.48 (4.78)	70.28 (5.10)
RRC	66.26 (2.99)	82.17 (5.08)	78.50 (2.35)	70.98 (3.48)	73.95 (3.46)	67.83 (4.80)

Table E.1.: ***Results of different illumination normalization methods for the ChimpZoo dataset.*** The table shows the obtained accuracies and standard deviations of the experiments conducted in section 5.4.1 for the ChimpZoo dataset. The best results are printed in boldface and were achieved by the proposed global face recognition pipeline as part of the Primate Recognition Framework (PRF).

Acc (Std.) [%]	NONE	HIST	CLAHE	MSR	IMADJUST	GAMMA + DOG
PRF (global)	**81.68 (2.54)**	**80.87 (1.87)**	**80.77 (2.28)**	**79.66 (2.21)**	**81.17 (3.05)**	**79.66 (3.79)**
Eigenfaces	47.57 (3.43)	56.07 (3.19)	54.15 (2.28)	46.46 (3.58)	50.61 (2.66)	37.15 (3.59)
Fisherfaces	6.07 (1.73)	42.71 (3.39)	44.53 (1.28)	44.74 (1.55)	23.68 (1.96)	19.64 (1.75)
Laplacianfaces	56.28 (3.28)	62.25 (3.16)	63.16 (2.87)	62.75 (4.41)	57.39 (1.79)	54.45 (2.10)
Randomfaces	52.53 (1.02)	58.70 (3.89)	51.92 (3.30)	51.72 (4.48)	53.64 (0.88)	40.99 (0.79)
BDCT	58.50 (2.36)	58.70 (3.38)	55.97 (3.12)	57.29 (2.77)	58.30 (2.70)	53.54 (1.95)
GSRC	68.83 (2.57)	75.40 (2.31)	74.60 (2.16)	71.56 (3.23)	70.24 (2.60)	70.85 (4.98)
RSC	58.70 (2.27)	65.69 (4.78)	64.37 (3.03)	62.25 (4.67)	59.31 (2.35)	49.70 (1.37)
RRC	54.05 (2.96)	63.06 (3.52)	58.91 (2.03)	54.76 (4.81)	56.38 (2.77)	45.95 (2.44)

Table E.2.: ***Results of different illumination normalization methods for the ChimpTaï dataset.*** The table shows the obtained accuracies and standard deviations of the experiments conducted in section 5.4.1 for the ChimpTaï dataset. The best results are printed in boldface and were achieved by the proposed global face recognition pipeline as part of the PRF.

Appendix F.

Results for Synthetically Generated Train and Test Data

Spotlight Position						
Acc (Std.) [%]	**Train Data**	**ExtremeLeft**	**MidLeft**	**Frontal**	**MidRight**	**ExtremeRight**
PRF (global)	real	25.02 (6.25)	44.80 (8.49)	55.99 (7.19)	49.19 (6.49)	24.45 (6.24)
	synth.	**84.25 (5.07)**	**82.68 (5.43)**	**77.46 (5.80)**	**82.56 (5.50)**	**84.47 (3.07)**
GSRC	real.	12.93 (4.00)	35.38 (7.13)	52.16 (8.98)	45.97 (8.03)	8.90 (2.46)
	synth.	**81.12 (5.73)**	**76.92 (4.44)**	**69.28 (7.82)**	**77.70 (6.59)**	**80.65 (5.16)**
Laplacianfaces	real.	28.94 (7.21)	45.13 (5.23)	51.84 (6.25)	46.39 (8.29)	17.12 (3.90)
	synth.	**76.56 (6.62)**	**73.95 (3.51)**	**69.45 (6.01)**	**74.86 (5.28)**	**75.01 (5.69)**
RSC	real	9.78 (3.99)	25.71 (4.58)	52.19 (6.44)	31.82 (3.14)	6.64 (3.15)
	synth.	**71.64 (6.25)**	**72.58 (4.54)**	**66.68 (6.07)**	**73.32 (5.91)**	**78.03 (3.51)**
RRC	real	7.52 (2.33)	22.41 (4.94)	47.31 (10.29)	31.36 (6.10)	4.72 (1.43)
	synth.	**82.70 (5.64)**	**61.23 (7.05)**	**53.03 (6.94)**	**61.26 (5.31)**	**76.96 (8.32)**

Table F.1.: ***Accuracies and standard deviations for synthetic training and test data applied to spotlight position variation.*** The table compares the accuracies and standard deviations of five different face recognition algorithms trained on real and synthetic data respectively. Using synthetic data helps increasing the generalization capability of all algorithms towards different spotlight positions. The best accuracies are printed in boldface letters.

Spotlight Exposure				
Acc (Std.) [%]	**Train Data**	**ExtremeLow**	**Mid**	**ExtremeHigh**
PRF (global)	real	88.68 (5.17)	55.99 (7.19)	22.34 (2.80)
	synth.	**90.34 (5.18)**	**77.46 (5.80)**	**48.42 (6.35)**
GSRC	real.	78.55 (5.68)	52.16 (8.98)	13.47 (4.10)
	synth.	**84.48 (5.58)**	**69.28 (7.82)**	**29.25 (8.85)**
Laplacianfaces	real.	67.22 (6.36)	51.84 (6.25)	19.98 (6.61)
	synth.	**77.47 (5.31)**	**69.45 (6.01)**	**37.82 (6.69)**
RSC	real	68.78 (6.08)	52.19 (6.44)	19.99 (7.07)
	synth.	**79.40 (5.46)**	**66.68 (6.07)**	**31.68 (6.01)**
RRC	real	62.73 (9.99)	47.31 (10.29)	15.23 (5.43)
	synth.	**81.64 (4.38)**	**53.03 (6.94)**	**24.89 (6.37)**

Table F.2.: ***Accuracies and standard deviations for synthetic training and test data applied to spotlight exposure variation.*** The table compares the accuracies and standard deviations of five different face recognition algorithms trained on real and synthetic data respectively. Using synthetic data helps increasing the generalization capability of all algorithms towards different spotlight exposure levels. The best accuracies are printed in boldface letters. However, overexposure still seems to be a problem for all algorithms due to a "white-out" effect of the facial features.

Acc (Std.) [%]	Train Data	-30	-20	-10	0	10	20	30
Horizontal Pose Variation								
PRF (global)	real	61.23 (7.45)	82.92 (4.58)	90.59 (4.28)	**94.76 (2.82)**	89.38 (5.56)	83.65 (5.53)	66.60 (4.83)
	synth.	**91.16 (4.37)**	**91.50 (4.81)**	**94.35 (2.85)**	93.97 (3.75)	**94.22 (2.48)**	**93.72 (3.48)**	**90.21 (2.83)**
GSRC	real	28.80 (6.14)	54.10 (5.60)	77.14 (4.32)	**87.76 (3.03)**	75.25 (5.82)	60.86 (5.27)	40.43 (5.43)
	synth.	**83.73 (5.21)**	**87.22 (4.69)**	**87.43 (3.83)**	87.60 (4.00)	**86.20 (5.52)**	**86.40 (4.12)**	**85.32 (3.71)**
Laplacianfaces	real	23.83 (5.04)	44.47 (6.53)	61.59 (6.96)	**87.24 (4.15)**	66.71 (7.95)	51.63 (4.18)	36.70 (3.00)
	synth.	**76.61 (5.82)**	**80.79 (3.13)**	**81.30 (3.36)**	82.54 (4.84)	**82.72 (5.54)**	**81.32 (4.50)**	**77.11 (5.81)**
RSC	real	29.80 (6.61)	42.86 (4.22)	63.03 (7.96)	83.39 (2.83)	67.34 (5.46)	51.95 (4.80)	37.39 (3.94)
	synth.	**81.34 (4.73)**	**84.29 (4.44)**	**84.46 (5.71)**	**83.77 (5.67)**	**83.26 (6.20)**	**83.60 (5.72)**	**80.96 (4.08)**
RRC	real	27.82 (4.62)	43.04 (5.08)	59.09 (8.26)	**84.09 (4.76)**	64.00 (5.76)	47.91 (6.81)	35.49 (4.27)
	synth.	**80.24 (5.77)**	**82.14 (6.03)**	**81.94 (6.13)**	80.41 (4.56)	**83.57 (4.82)**	**82.50 (4.04)**	**79.89 (4.88)**
Vertical Pose Variation								
PRF (global)	real	64.57 (7.49)	78.26 (7.98)	87.82 (4.51)	**94.76 (2.82)**	90.22 (5.31)	85.13 (4.06)	47.41 (7.70)
	synth.	**92.35 (3.61)**	**93.38 (3.21)**	**93.55 (2.82)**	93.97 (3.75)	**93.72 (4.31)**	**92.50 (3.79)**	**88.47 (5.81)**
GSRC	real	47.44 (4.90)	67.90 (6.03)	77.35 (4.58)	**87.76 (3.03)**	78.69 (4.45)	66.77 (2.75)	31.10 (5.36)
	synth.	**84.65 (5.55)**	**87.41 (4.62)**	**87.43 (4.43)**	87.60 (4.00)	**84.45 (3.46)**	**84.97 (4.57)**	**84.44 (3.99)**
Laplacianfaces	real	34.99 (4.30)	45.33 (5.65)	62.17 (8.13)	**87.24 (4.15)**	64.58 (6.76)	45.32 (3.13)	21.38 (6.23)
	synth.	**79.21 (4.98)**	**80.44 (4.49)**	**80.77 (4.59)**	82.54 (4.84)	**80.42 (4.81)**	**78.36 (6.36)**	**74.63 (4.49)**
RSC	real	37.24 (3.94)	48.82 (5.70)	64.21 (4.49)	83.39 (2.83)	66.28 (7.00)	46.69 (5.85)	16.60 (4.81)
	synth.	**80.10 (4.39)**	**84.10 (4.31)**	**84.62 (5.66)**	**83.77 (5.67)**	**83.04 (4.46)**	**80.42 (4.46)**	**76.79 (4.03)**
RRC	real	32.43 (5.52)	47.61 (6.34)	58.60 (6.09)	**84.09 (4.76)**	61.73 (6.54)	39.51 (2.88)	17.49 (4.54)
	synth.	**82.67 (3.85)**	**83.19 (5.06)**	**84.26 (4.82)**	80.41 (4.56)	**83.88 (6.00)**	**82.16 (4.73)**	**79.56 (4.76)**

Table F.3.: ***Accuracies and standard deviations for synthetic training and test data applied to pose variation.*** The table compares the accuracies and standard deviations of five different face recognition algorithms trained on real and synthetic data respectively. Using synthetic data helps increasing the generalization capability of all algorithms towards pose variations. The best accuracies are printed in boldface letters.

Appendix G.

Copyrights of Photographs

[I01] © Dr. Hjalmar Kühl – MPI EVA (2014), Taï Nationalpark
[I02] © Dr. Tobias Deschner – MPI EVA (1998/1999), Taï Nationalpark
[I03] © Dr. Damien Caillaud – MPI EVA (2004), Odzala Nationalpark
[I04] © Laura Aporius – MPI EVA (2010), WKPRC (Zoo Leipzig)
[I05] © Roz Sandwell and Tilo Burghardt – University of Bristol (2013)
[I06] © Mitteldeutscher Rundfunk - Thüringen Journal (2012)
[I07] © Mitteldeutscher Rundfunk - Sachsenspiegel (2012)

List of Figures

List of Tables

List of Acronyms

k-NN k-Nearest-Neighbor. 22–24, 30, 158, 217

AAM Active Appearance Model. 125

ASM Active Shape Model. 125

AUC Area Under the Curve. 156, 178, 179, 187

BRIEF Binary Robust Independent Elementary Features. 173

CCA Canonical Correlation Analysis. 121

ChimpZoo The dataset of captured chimpanzee individuals gathered at the Zoo of Leipzig, Germany. 145, 146, 148–153, 159–161, 164, 166, 168, 173–181, 183–188, 209, 211, 221

CLAHE Contrast Limited Adaptive Histogram Equalization. 162, 164

CS Compressed Sensing. 7, 25, 26, 30, 93, 113

DAISY A Fast Local Descriptor for Dense Matching. 172–174

DCT Discrete Cosine Transform. 87, 94, 95, 159, 165

DoG Difference of Gaussian. 163

DPM Deformable Part Based Models. 105

EBGM Elastic Bunch Graph Matching. 97, 98

EER Equal Error Rate. 156, 178, 179, 187

ELGTPHS Enhanced Local Gabor Ternary Pattern Histogram Sequence. 6, 7, 108, 109, 111, 112, 117, 126, 127, 161, 164, 167, 201, 204

ELTP Extended Local Ternary Patterns. 7, 108–111, 116, 126, 165, 227

EVD Extreme Value Distribution. 134–138, 193

fps frames per second. 182

FRVT Face Recognition Vendor Test. 90

GMP Gabor Magnitude Picture. 12, 96, 108–111, 116, 126, 218, 227, 228

GorillaZoo The dataset of captured gorilla individuals gathered at the Zoo of Leipzig, Germany. 145, 146, 148–150, 181, 182, 219

GPSR Gradient Projection for Sparse Reconstruction. 114

KCCA Kernel Canonical Correlation Analysis. 121

LBP Local Binary Patterns. 10, 12–16, 29, 63, 64, 78, 87, 95, 96, 111, 217

LDA Linear Discriminant Analysis. 19–21, 29, 61, 76, 86, 91, 92, 112, 158, 161

LFW Labeled Faces in the Wild. 90

LGBPHS Local Gabor Binary Pattern Histogram Sequence. 96

LIBSVM Library for Support Vector Machines. 119, 121, 133

LPP Locality Preserving Projections. 7, 19, 21, 22, 29, 92, 107, 108, 111–113, 116, 118, 120, 158, 161, 173, 193, 228, 229

LTP Local Ternary Patterns. 10, 15, 16, 95, 111, 217

MATLAB® MATrix LABoratory. A multi-paradigm numerical computing environment.. 162

MLE Maximum Likelihood Estimation. 7, 93, 135, 137, 227

MSR Multi Scale Retinex. 162, 164

ORB Oriented FAST and Rotated BRIEF. 173, 174

PCA Principal Component Analysis. 19–21, 29, 73, 86, 91, 92, 94, 97, 112, 113, 118, 126, 158, 159, 161

PDF Probability Density Function. 93, 135, 137, 138, 184

PRF Primate Recognition Framework. v, vii, 1, 2, 4–10, 29, 31, 88, 99, 100, 102, 107, 111, 112, 114, 117, 121, 124, 129, 138, 141, 145, 146, 150, 152–154, 157, 159–162, 166, 167, 169, 171, 174–178, 181, 182, 184, 187, 189, 199–202, 205, 206, 209, 211, 213, 218, 219, 221, 229

RBF Radial Basis Function. 28, 98, 108, 119, 120, 126, 227

RIP Restricted Isometry Property. 93

ROC Receiver Operating Characteristic. 156, 178–180, 186, 194, 219

RRC Regularized Robust Coding. 93, 94, 158, 159, 166, 169, 171, 213

RSC Robust Sparse Coding. 93, 94, 158, 159, 166, 169, 171, 213

SCI Sparse Concentration Index. 133

SHORE™ Sophisticated High-Speed Object Recognition Engine. 9, 103–106, 117, 118, 123, 124, 130, 139, 152, 153, 173, 181–183, 185, 188, 195–197, 200, 202, 218

SIFT Scale Invariant Feature Transform. 16–18, 29, 172–174, 204, 217

SRC Sparse Representation Classification. 7, 25, 26, 30, 93, 94, 108, 113–116, 121, 122, 132–134, 158, 159, 193

SURF Speeded-Up Robust Features. 18, 19, 29, 108, 117, 118, 120, 127, 172–174, 193, 204, 217, 218

SVM Support Vector Machine. 26–28, 30, 108, 119–122, 126, 133, 134, 173, 193, 227, 229

List of Symbols

C A penalty parameter for SVM classification. A large C corresponds to high penalties for classification errors.. 27, 28, 119, 120, 126, 173

$\boldsymbol{\delta}_i$ The characteristic function of class i which is 1 for samples if class i and 0 elsewhere.. 26, 114, 132, 133, 228

$\boldsymbol{A}$ Normalized matrix of training samples of size $m \times l$ that holds the normalized feature vectors projected into the low dimensional subspace.. 25, 26, 113–116, 132, 133, 228

α The slope of the power spectrum of an image. 127, 128

B The number of blocks the GMPs are divided before Extended Local Ternary Patterns (ELTP) feature histograms are extracted.. 111, 116

$\boldsymbol{S}_\mathrm{b}$ The between-class scatter matrix $\boldsymbol{X}$.. 20, 21

γ A user-defined parameter used for creating an Radial Basis Function (RBF) kernel for non-linear Support Vector Machine (SVM) classification.. 119, 120, 126, 173

C Number of classes in the dataset. 20, 21, 25, 114, 121, 133, 138, 154, 155, 158, 159, 178, 193, 227, 229

c The class of a given sample. 20, 21

$\boldsymbol{c}$ The class vector of size $C \times 1$. 27, 113

$\boldsymbol{S}_\mathrm{PCA}$ The covariance matrix of the data samples $\boldsymbol{X}$.. 20

$\boldsymbol{s}_c$ The cumulative score vector of size $C \times 1$ used for the proposed frame weighting approach.. 138

$\boldsymbol{D}$ The diagonal matrix of size $N \times N$ whose entries are column (or row since $\boldsymbol{S}$ is symmetric) sums of $\boldsymbol{S}$.. 22, 228

$\mathcal{L}$ The Log-Likelihood-Function that have to be maximized for Maximum Likelihood Estimation (MLE).. 136

d_M The Mahalanobis-distance from test vector $\boldsymbol{t}$ to cluster of observations $\boldsymbol{O}$.. 135

Δd_M The difference of Mahalanobis-distances from the test sample $\boldsymbol{t}$ the cluster of correct and incorrect classifications.. 134, 135, 137, 138

F The user-defined number of the rotations for the Gabor kernels.. 10, 110, 116, 126, 228

$\boldsymbol{\mathcal{G}}$ Adjacency graph of Locality Preserving Projections (LPP) to model the local neighborhood of the feature points. 22, 112, 113

$\boldsymbol{g}^{(\rho)}$ The downscaled and vectorized version of the GMP. 126

$\boldsymbol{L}$ The Laplacian matrix of size $N \times N$. $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$. 22, 228

l Number of training samples.. 25, 26, 114, 193, 227, 228

m Size of the low-dimensional feature space after transformation. 19–25, 27, 28, 92, 112–114, 116, 120, 126, 159, 173, 193, 227, 229

μ The index of the rotation for the Gabor kernels. $\mu = 0, \cdots, F-1$. 10–12, 109, 110, 126, 228

ν The index of the scales for the Gabor kernels. $\nu = 0, \cdots, S-1$. 10–12, 109, 110, 126, 228

N Number of image in the data set. 19, 20, 22, 112, 227–229

n Size of the original feature space. 19–21, 92, 112, 113, 229

$\boldsymbol{n}$ The normal vector of a (hyper-)plane. 27

$\boldsymbol{G}_{\mu,\nu}(z)$ The Gabor wavelet transform resulting from the convolution of an input image $I(z)$ with the complex Gabor kernel $\boldsymbol{\Psi}_{\mu,\nu}(z)$.. 11, 12, 109, 110, 126, 228

$\boldsymbol{p}$ The sparse coefficient vector of size $\mathbb{R}^{l\times 1}$.. 25, 26, 113–115, 132, 133, 228

$|\boldsymbol{G}_{\mu,\nu}(z)|$ The Gabor magnitude picture. 12, 110, 126

$\boldsymbol{\Phi}_{\mu,\nu}(z)$ The phase of a Gabor kernel. 12

$\boldsymbol{\Psi}_{\mu,\nu}(z)$ The Gabor kernel. 10, 11, 109, 110, 228

r_i The residual of class i: $r_i(\boldsymbol{t}) = \|\boldsymbol{t} - \boldsymbol{A}(\boldsymbol{\delta}_i \odot \hat{\boldsymbol{p}}_1)\|_2$.. 26, 114, 228

ρ The scaling factor for downsampling the GMP. 126, 228

S The user-defined number of the scales for the Gabor kernels.. 110, 116, 126, 228

$\boldsymbol{S}$ The slope of the power spectrum of an image. 127, 128

$\boldsymbol{S}$ Sparse symmetric matrix of size $N \times N$ that holds the weights of the edge joint vertices for LPP. 22, 113, 118, 227, 228

$\boldsymbol{s}$ The score vector. 121, 134, 138, 154, 156

$\boldsymbol{v}$ The confidence vector of size $\mathbb{R}^{5\times 1}$ which comprises five different confidence measures of the proposed PRF.. 133, 134

$\boldsymbol{W}$ Unitary transformation matrix of size $n \times m$ to project a high dimensional feature vector into a smaller dimensional subspace. 19, 20, 22, 112, 113, 229

$\wp$ The rank of a classification. $\wp : \{1, \cdots, C\}$. 121, 154, 155, 229

$\boldsymbol{w}$ The weight vector for the proposed decision fusion scheme to combine the results of global and local features.. 121

$\boldsymbol{w}$ Column vectors of the unitary transformation matrix $\boldsymbol{W}$. 20–22

$\boldsymbol{S}_{\mathrm{w}}$ The within-class scatter matrix $\boldsymbol{X}$.. 20, 21, 113

$\boldsymbol{X}$ Matrix of size $n \times N$ that holds the feature vectors $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$. 19, 20, 22, 112, 227, 229

$\boldsymbol{x}$ Feature vector in original feature space. 19–22, 112, 113, 118, 126, 229

ξ A slack variable introduced to handle non-separable data for SVM classification. $\xi \geq 0$. 27, 28, 229

$\boldsymbol{t}$ Projected feature vector of a test sample. 23–27, 113–115, 119, 132–135, 138, 193, 228

$\boldsymbol{y}$ Projected feature vector in low dimensional subspace. 19, 22–24, 27, 28, 112–114, 116, 119

z The pixel location $z = (x, y)$. 10–13, 15, 109–111, 126, 228, 229

Bibliography

Publications as Co- or First Author

[1] A. Loos, H. Kühl, A. Ernst, O. Wagner, H. Lukashevich, T. Burghardt, J. Garbas, and C. Li, "SAISBECO - A Semi-Automated Audiovisual Species and Individual Identification System for Behavioral Ecological Research and Conservation: Final Report," tech. rep., Fraunhofer IDMT, Fraunhofer IIS, MPI EVAN, University of Bristol, 2013.

[2] A. Loos and H. Kuehl, "SAISBECO - A Semi-Automated Audiovisual Species and Individual Identification System for Behavioral Ecological Research and Conservation," in *International Scientific Colloquium (IWK)*, (Ilmenau, Germany), 2010.

[3] A. Loos, M. Pfitzer, and L. Aporius, "Facial Recognition for Primate Photo Identification," in *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, vol. 54, (Delft, Netherlands), 2011.

[4] A. Loos, M. Pfitzer, and L. Aporius, "Identification of Great Apes Using Face Recognition," in *European Signal Processing Conference (EUSIPCO)*, (Barcelona, Spain), 2011.

[5] A. Loos and M. Pfitzer, "Towards Automated Visual Identification Of Primates Using Face Recognition," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*, (Vienna, Austria), 2012.

[6] A. Loos, "Identification of Great Apes Using Gabor Features and Locality Preserving Projections," in *ACM International Workshop on Multimedia Analysis for Ecological Data (MAED) in conjunction with ACM Multimedia*, (Nara, Japan), 2012.

[7] A. Loos and A. Ernst, "Detection and Identification of Chimpanzee Faces in the Wild," in *IEEE International Symposium on Multimedia (ISM)*, (Irvine, California, USA), 2012.

[8] A. Loos, "Chimpanzee Identification Using Global and Local Features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Vancouver, Canada), 2013.

[9] A. Loos and A. Ernst, "An Automated Chimpanzee Identification System Using Face Detection and Recognition," *EURASIP Journal on Image and Video Processing: Special Issue on Animal Behaviour Understanding in Image Sequences*, vol. 2013, no. 49, 2013.

[10] R. Sandwell, A. Loos, and T. Burghardt, "Synthesising Unseen Image Conditions to Enhance Classification Accuracy for Sparse Datasets: Applied to Chimpanzee Face Recognition," in *UK Computer Vision Student Workshop (BMVW at BMVC)*, 2013.

[11] A. Loos and T. A. M. Kalyanasundaram, "Face Recognition for Great Apes: Identification of Primates in Videos," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Brisbane, Australia), 2015.

[12] A. Loos, R. Paduschek, D. Kormann, and P. Dunker, "Evaluation of Algorithms for the Summarization of Photo Collections," in *Theseus/ImageCLEF Workshop on Visual Information Retrieval Evaluation*, (Corfu, Greece), 2009.

References by Other Authors

[13] J. Vié, C. Hilton-Taylor, S. Stuart, I.-T. W. C. Union, and I. S. S. Commission, *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species*. IUCN, 2009.

[14] P. D. Walsh, K. A. A. Abernethy, and M. Bermejo, "Catastrophic Ape Decline in Western Equatorial Africa," *Nature*, vol. 422, pp. 611–614, 2003.

[15] G. Campbell, H. Kühl, P. Kouame, and C. Boesch, "Alarming Decline of West African Chimpanzees in Côte d'Ivoire," *Current Biology*, vol. 18, no. 19, pp. R904–R905, 2008.

[16] R. F. W. Barnes and K. L. Jensen, "How To Count Elephants in Forests," *IUCN African Elephant and Rhino Specialist Group Technical Bulletin*, vol. 1, pp. 1–6, 1987.

[17] S. Buckland, D. Anderson, K. Burnham, J. Laake, D. Borchers, and L. Thomas, *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, 2001.

[18] J. M. Rowcliffe and C. Carbone, "Surveys using Camera Traps: Are We Looking to a Brighter Future?," *Animal Conservation*, vol. 11, no. 3, pp. 185–186, 2008.

[19] P. Kappeler and C. van Schaik, "Evolution of Primate Social Systems," *International Journal of Primatology*, vol. 23, pp. 707–740, 2002.

[20] D. Houle, D. Govindaraju, and S. Omholt, "Phenomics: The Next Challenge," *Nature Reviews Genetics*, vol. 11, pp. 855–866, 2010.

[21] T. Freeberg, R. Dunbar, and T. Ord, "Social Complexity as a Proximate and Ultimate Factor in Communicative Complexity," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 367, pp. 1785–1801, 2012.

[22] C. Boesch, K. Brandenburg, and H. Gerhäuser, "SAISBECO: A Semi-Automated Audiovisual Species and Individual Identification System for Behavioral Ecological Research and Conservation," Project Proposal, Fraunhofer IDMT, Fraunhofer IIS, and Max-Planck-Institute EVAN, 2010.

[23] J. A. Royle and J. D. Nichols, "Estimating Abundance from Repeated Presence-Absence Data or Point Counts," *Ecology*, vol. 84, pp. 777–790, 2003.

[24] M. Efford, D. Borchers, and A. Byrom, "Density Estimation by Spatially Explicit Capture-Recapture: Likelihood-Based Methods," *Modeling Demographic Processes In Marked Populations Environmental and Ecological Statistics*, vol. 3, pp. 255–269, 2009.

[25] D. Borchers and M. Efford, "Spatially Explicit Maximum Likelihood Methods for Capture-Recapture Studies," *Biometrics*, vol. 64, no. 2, pp. 377–385, 2008.

[26] M. Lindberg, "A Review of Designs for Capture-Mark-Recapture Studies in Discrete Time," *Journal of Ornithology*, vol. 152, no. 2, pp. 355–370, 2012.

[27] M. Nietfeld, M. Barrett, and N. Silvy, "Wildlife Marking Techniques," in *Research and Management Techniques for Wildlife Habitats* (T. W. Society, ed.), pp. 140–168, 1994.

[28] B. Williams, J. Nichols, and M. Conroy, *Analysis and Management of Animal Populations: Modeling, Estimation, and Decision Making.* Academic Press, 2002.

[29] D. Murray and M. Fuller, "A Critical Review of the Effects of Marking on the Biology of Vertebrates," *Research Techniques in Animal Ecology: Controversies and Consequences*, pp. 15–64, 2000.

[30] R. Powell and G. Proulx, "Trapping and Marking Terrestrial Mammals for Research: Integrating Ethics, Performance Criteria, Techniques, and Common Sense," *ILAR Journal*, vol. 44, pp. 259–276, 2003.

[31] M. McCarthy and K. Parris, "Clarifying the Effect of Toe Clipping on Frogs with Bayesian Statistics," *Journal of Applied Ecology*, vol. 41, pp. 780–786, 2004.

[32] A. O'Connell, J. Nichols, and K. Karanth, *Camera Traps in Animal Ecology: Methods and Analyses.* Springer, 2011.

[33] J. Head, C. Boesch, M. Robbins, L. Rabal, L. Makaga, and H. Kühl, "Effective Socio-Demographic Population Assessment of Elusive Species for Ecology and Conservation Management," *Ecology and Evolution*, vol. 3, no. 9, pp. 2903–2916, 2013.

[34] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface.* John Wiley & Sons Inc., 2002.

[35] P. Aichroth, J. Björklund, F. Stegmaier, T. Kurz, and G. Miller, "State of the Art in Cross-Media Analysis, Metadata Publishing, Querying and Recommendations," tech. rep., Fraunhofer Institute for Digital Media Technology (IDMT), 2014.

[36] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1615–1630, Oct. 2005.

[37] D. Gabor, "Theory of Communication," *Journal of the Institution of Eletrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–457, 1946.

[38] T. Ojala, M. Pietikäinen, and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Feature Distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[39] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[40] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[41] J. Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *Journal of the Optical Society of America (JOSA) A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[42] T. Lee, "Image Representation Using 2D Gabor Wavelets," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 1–13, 1996.

[43] T. Weldon, W. Higgins, and D. Dunn, "Efficient Gabor Filter Design for Texture Segmentation," *Pattern Recognition*, vol. 29, no. 12, pp. 2005–2015, 1996.

[44] J. Shao and W. Förstner, "Gabor Wavelets for Textrue Edge Extraction," in *Proc. SPIE 2357, ISPRS Commission III Symposium: Spatial Information from Digital Photogrammetry and Computer Vision*, 1994.

[45] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 19, no. 7, pp. 775–779, 1997.

[46] C. Liu and H.Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[47] S. Xie, S. Shan, and C. X., "Fusing Local Patterns of Gabor Magnitude and Phase for Face Recognition," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1349–1361, 2010.

[48] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *European Conference on Computer Vision (ECCV)*, 2010.

[49] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, "Local Binary Patterns for Still Images," in *Computer Vision Using Local Binary Patterns*, vol. 40 of *Computational Imaging and Vision*, ch. 2, pp. 13—-47, London: Springer London, computatio ed., 2011.

[50] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 12, pp. 2037–2041, 2006.

[51] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[52] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[53] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, pp. 585–591, 2001.

[54] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[55] M. Brand, "Charting a Manifold," *Advances in Neural Information Processing Systems (NIPS)*, vol. 16, pp. 961–968, 2003.

[56] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[57] J. Tenenbaum, V. de Silva, and J. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 12, pp. 2319–2323, 2000.

[58] X. He, *Locality Preserving Projections*. PhD thesis, University of Chicago, 2005.

[59] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, second edition ed., 1990.

[60] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[61] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal on Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[62] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis andPattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[63] Y. Chang, C. Hu, and M. Turk, "Manifold of Facial Expression," in *IEEE International Workshop Analysis and Modeling of Faces and Gestures*, 2003.

[64] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[65] H. Seung and D. Lee, "The Manifold Ways of Perception," *Science*, vol. 290, no. 5500, pp. 2268–2269, 2000.

[66] A. Shashua, A. Levin, and S. Avidan, "Manifold Pursuit: A New Approach to Appearance Based Recognition," in *International Conference on Pattern Recognition (ICPR)*, 2002.

[67] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[68] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[69] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 328–40, Mar. 2005.

[70] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer New York, 2009.

[71] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, pp. 210–27, Feb. 2009.

[72] E. Amaldi and V. Kann, "On the Approximability of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems," *Theoretical Computer Science*, vol. 209, pp. 237–260, 1998.

[73] D. Donoho, "Compressed Sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[74] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.

[75] V. Vapnik, *Statistical Learning Theory.* John Wiley and Sons, Inc., New York, 1998.

[76] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[77] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[78] C.-w. Hsu, C.-c. Chang, and C.-j. Lin, "A Practical Guide to Support Vector Classification," tech. rep., Department of Computer Science, National Taiwan University, Taiwan, 2010.

[79] B. Boser, I. Guyon, and V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *5th Annual Workshop on Computational Learning Theory*, (Pittsburgh, PA,USA), pp. 144–152, 1992.

[80] A. Aizerman, E. Braverman, and L. Rozoner, "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[81] S. Baban, "SVM with Polynomial Kernel Visualization." online: `http://www.sbaban.org/2013/11/05/svm-with-polynomial-kernel-visualization/`. Last visit: April 24^{th}, 2014.

[82] K. B. Duan and S. S. Keerthi, "Which Is the Best Multiclass SVM Method? An Empirical Study," *Multiple Classifier Systems. Lecture Notes in Computer Science*, vol. 3541, pp. 278–285, 2005.

[83] C.-W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[84] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[85] H. S. Kühl and T. Burghardt, "Animal Biometrics: Quantifying and Detecting Phenotypic Appearance," *Trends in Ecology & Evolution*, vol. 28, pp. 432–441, Mar. 2013.

[86] M. Lopes, A. Koerich, N. Silla, and C. Kaestner, "Feature Set Comparison for Automatic Bird Species Identification," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (Anchorage, AK, USA), pp. 965–970, 2011.

[87] S.-G. Huang, X.-L. Li, M.-Q. Zhou, and G.-H. Geng, "SURF-Based Multi-scale Resolution Histogram for Insect Recognition," in *International Conference on Artificial Intelligence and Computational Intelligence (ICAI)*, no. 1, (Las Vegas, Nevada, USA), pp. 445–448, Ieee, 2009.

[88] I. Potamitis, T. Ganchev, and N. Fakotakis, "Automatic Acoustic Identification of Insects Inspired by the Speaker Recognition Paradigm," in *Interspeech*, (Pittsburgh, PA, USA), 2006.

[89] A. Kershenbaum, L. Sayigh, and V. Janik, "The Encoding of Individual Identity in Dolphin Signature Whistles: How Much Information is Needed?," *PLoS ONE*, vol. 8, no. 10, 2013.

[90] M. P. Huijser, T. D. Holland, M. Blank, M. C. Greenwood, P. T. Mcgowen, B. Hubbard, and S. Wang, "The Comparison of Animal Detection Systems in a Test-Bed: A Quantitative Comparison of System Reliability and Experiences with Operation and Maintenance Final report," tech. rep., Federal Highway Administration, Helena, Montana, USA, 2009.

[91] U. Schatzmann, "Kennzeichnungsmethoden bei Pferden mittels Heissbrand und Transponder-Implantation unter besonderer Berücksichtigung von Schmerzen und Leiden," tech. rep., University of Bern, 2012.

[92] D. Edwards, A. Johnston, and D. Pfeiffer, "A Comparison of Commonly Used Ear Tags on the Ear Damage of Sheep," *Animal Welfare*, vol. 10, pp. 141–151, 2001.

[93] T. Davis and K. Ovaska, "Individual Recognition of Amphibians: Effects of Toe Clipping and Fluorescent Tagging on the Salamander Plethodon Vehiculum," *Journal of Herpetology*, vol. 35, no. 2, pp. 217–225, 2001.

[94] C. Schmid, "Constructing Models for Content-Based Image Retrieval," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, (Kauai, Hawaii, USA), pp. 39–45, 2001.

[95] T. L. Berg and D. A. Forsyth, "Animals on the Web," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (New York, USA), pp. 1463–1470, 2006.

[96] G. Cox, "Template Matching and Measures of Match in Image Processing," tech. rep., Department of Electrical Engineering, University of Cape Town, 1995.

[97] G. Kastberger, M. Maurer, F. Weihmann, M. Ruether, T. Hoetzl, I. Kranner, and H. Bischof, "Stereoscopic Motion Analysis in Densely Packed Clusters: 3D Analysis of the Shimmering Behaviour in Giant Honey Bees.," *Frontiers in Zoology*, vol. 8, pp. 2–18, Jan. 2011.

[98] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[99] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Kauai, Hawai, USA), pp. 511–518, 2001.

[100] Y. Freund and R. Schapire, "A Short Introduction to Boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 15, no. 5, pp. 771–780, 1999.

[101] E. Hjelmas and B. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001.

[102] M. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[103] S. Roy, S. Roy, and S. Bandyopadhyay, "A Tutorial Review on Face Detection," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 8, 2012.

[104] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Computer Vision and Pattern Recognition*, 2005.

[105] T. Burghardt, J. Calic, and B. T. Thomas, "Tracking Animals in Wildlife Videos Using Face Detection," in *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, (London, UK), 2004.

[106] T. Burghardt and J. Calic, "Real-Time Face Detection and Tracking of Animals," in *Seminar on Neural Network Applications in Electrical Engineering (NEUREL)*, (Belgrade, Serbia), pp. 27–32, 2006.

[107] J. Shi and C. Tomasi, "Good Features to Track," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, Washington, USA), pp. 593–600, 1994.

[108] W. Zhang, J. Sun, and X. Tang, "Cat Head Detection - How to Effectively Exploit Shape and Texture Features," in *European Conference on Computer Vision (ECCV)*, (Marseille, France), pp. 802–816, Springer, 2008.

[109] W. Zhang, J. Sun, and X. Tang, "From Tiger to Panda: Animal Head Detection," *IEEE Transactions on Image Processing*, vol. 20, pp. 1696–1708, June 2011.

[110] T. Kozakaya, S. Ito, S. Kubota, and O. Yamaguchi, "Cat Face Detection with Two Heterogeneous Features," in *International Conference on Image Processing (ICIP)*, (Cairo, Egypt), pp. 1213–1216, Ieee, Nov. 2009.

[111] R. Sumner and L. Ross, "Extension of the Viola-Jones Detector – Application to Cat Faces," tech. rep., Boston University, Boston, 2012.

[112] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection," *PSIVT, LNCS*, vol. 5414, pp. 37–47, 2009.

[113] A. Ernst and C. Küblbeck, "Fast Face Detection and Species Classification of African Great Apes," in *International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, (Klagenfurt, Austria), pp. 279–284, IEEE, Aug. 2011.

[114] R. Zabih and J. Woodfill, "Non-Parametric Local Transforms for Computing Visual Correspondence," in *European Conference on Computer Vision (ECCV)*, pp. 151–158, 1994.

[115] R. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-Rated Predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[116] B. Miranda, J. Salas, and P. Vera, "Bumblebees Detection and Tracking," in *Workshop on Visual Observation and Analysis of Animal and Insect Behavior (VAIB)*, 2012.

[117] S. Gu, Y. Zheng, and C. Tomasi, "Efficient Visual Object Tracking with Online Nearest Neighbor Classifier," in *Asian Conference on Computer Vision (ACCV)*, 2010.

[118] H. Stahl, K. Schädler, and E. Hartung, "Capturing 2D and 3D Biometric Data of Farm Animals under Real-Life Conditions," in *International Workshop on Computer Image Analysis in Agriculture*, no. 17, (Valencia, Spain), 2012.

[119] D. Ramanan and D. Forsyth, "Using Temporal Coherence to Build Models of Animals," in *International Conference on Computer Vision (ICCV)*, (Nice, France), pp. 1–8, 2003.

[120] D. Ramanan, D. A. Forsyth, and K. Barnard, "Detecting, Localizing and Recovering Kinematics of Textured Animals," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, (San Diego, California, USA), pp. 635–642, Ieee, 2005.

[121] D. Ramanan, D. A. Forsyth, and K. Barnard, "Building Models of Animals from Video.," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, pp. 1319–34, Aug. 2006.

[122] P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[123] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Chal- lenge 2010 (VOC2010) Results," tech. rep., 2010.

[124] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "The Truth About Cats and Dogs," in *International Conference on Computer Vision (ICCV)*, (Barcelona, Spain), 2011.

[125] Y. Boykov and R. Veksler, H.and Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[126] R. Sandwell and T. Burghardt, "Chimpanzee Face Detection: An Automated System for Images Captured From Natural Environments," in *International Conference on Behaviour, Physiology and Genetics of Wildlife*, (Berlin, Germany), 2013.

[127] C. Harris and M. Stephens, "A Combined corner and Edge Detector," in *Alvey Vision Conference*, (Manchester, UK), pp. 147–151, 1988.

[128] E. Larios, H. Deng, W. Zhang, M. Sarpola, J. Yuen, R. Paasch, A. Moldenke, D. Lytle, S. R. Correa, E. Mortensen, L. Shapiro, T. Dietterich, and F. V. Generation, "Automated Insect Identification through Concatenated Histograms of Local Appearance Features," in *IEEE Workshop on Applications of Computer Vision (WACV)*, (Austin, Texas, USA), 2007.

[129] P. de Zeeuw, E. Pauwels, E. Ranguelova, D. Buonantony, and S. Eckert, "Computer Assisted Photo Identification of Dermochelys Coriacea," in *Visual Observation and Analysis of Animal and Insect Behavior (VAIB)*, (Istanbul, Turkey), pp. 165–172, 2010.

[130] T. Burghardt, *Visual Animal Biometrics - Automatic Detection and Individual Identification by Coat Pattern*. Phd thesis, University of Bristol, 2008.

[131] T. Burghardt and N. Campbell, "Generic Phase Curl Localisation for Individual Identification of Turing-Patterned Animals," in *Workshop on Visual Observation and Analysis of Animal and Insect Behavior (VAIB)*, (Istanbul, Turkey), pp. 17–21, 2010.

[132] D. Walther, D. R. Edgington, and C. Koch, "Detection and Tracking of Objects in Underwater Video," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Washington, District Columbia, USA), pp. 544–549, 2004.

[133] D. R. Edgington, D. E. Cline, D. Davis, I. Kerkez, and J. Mariette, "Detection, Tracking and Classifying Animals in Underwater Video," in *OCEANS*, (Boston, Massachusetts, USA), pp. 1 – 5, 2006.

[134] C. Spampinato, S. Palazzo, B. Boom, J. van Ossenbruggen, I. Kavasidis, R. D. Salvo, F.-P. Lin, D. Giordano, L. Hardman, and R. B. Fisher, "Understanding Fish Behavior During Typhoon Events in Real-Life Underwater Environments," *Multimedia Tools and Applications*, vol. 68, no. 1, 2012.

[135] I. Kavasidis and S. Palazzo, "Quantitative Performance Analysis of Object Detection Algorithms on Underwater Video Footage," in *ACM International Workshop on Multimedia Analysis for Ecological Data (MAED)*, (Nara, Japan), pp. 57 – 60, 2012.

[136] L. Itti and C. Koch, "Target Detection using Saliency-Based Attention," in *Workshop on Search and Target Acquisition (NATO Unclassified)*, (Utrecht, The Netherlands), pp. 3.1–3.10, 1999.

[137] R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[138] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," tech. rep., Department of Computer Science, University of North Carolina, 2006.

[139] C. Schmid and R. Mohr, "Local Grayvalue Invariants for Image Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.

[140] O. Barnich and M. van Droogenbroeck, "ViBe: A Universal Background Subtraction Algorithm for Video Sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[141] F. Porikli, "Multiplicative Background-Foreground Estimation Under Uncontrolled Illumination Using Intrinsic Images," in *IEEE Workshop on Motion and Video Computing*, (Breckenridge, CO, USA), pp. 20–27, 2005.

[142] F. Porikli, T. O., and P. Meer, "Covariance Tracking Using Model Update Based on Lie Algebra," in *Computer Vision and Pattern Recognition (CVPR)*, (New York, NY, USA), pp. 728–735, 2006.

[143] F. Porikli, "Achieving Real-Time Object Detection and Tracking under Extreme Conditions," *Journal on Real-Time Image Processing*, vol. 1, no. 1, pp. 33–40, 2006.

[144] J. Wawerla, S. Marshall, G. Mori, K. Rothley, and P. Sabzmeydani, "BearCam: Automated Wildlife Monitoring at the Arctic Circle," *Machine Vision and Applications*, vol. 20, pp. 303–317, Apr. 2009.

[145] P. Sabzmeydani and G. Mori, "Detecting Pedestrians by Learning Shapelet Features," in *Computer Vision and Pattern Recognition (CVPR)*, (Minneapolis, MN, USA), pp. 1–8, 2007.

[146] D. Song, N. Qin, Y. Xu, C. Y. Kim, D. Luneau, and K. Goldberg, "System and Algorithms for an Autonomous Observatory Assisting the Search for the Ivory-Billed Woodpecker," in *International Conference on Automation Science and Engineering (CASE)*, (Shanghai, China), pp. 200–205, 2006.

[147] A. Elgammal, D. Harwood, and L. Davis, "Non-Parametric Model for Background Subtraction," in *European Conference on Computer Vision (ECCV)*, (Dublin, Ireland), pp. 751–767, 2000.

[148] Z. Khan, R. A. Herman, K. Wallen, and T. Balch, "An Outdoor 3-D Visual Tracking System for the Study of Spatial Navigation and Memory in Rhesus Monkeys," *Behavior Research Methods*, vol. 37, no. 3, pp. 453–463, 2005.

[149] P. Khorrami, J. Wang, and T. Huang, "Multiple Animal Species Detection Using Robust Principal Component Analysis and Large Displacement Optical Flow," in *Workshop on Visual Observation and Analysis of Animal and Insect Behavior (VAIB)*, (Tsukuba, Japan), 2012.

[150] E. Candes, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?," *Journal of the ACM*, vol. 58, no. 3, 2011.

[151] T. Brox and J. Malik, "Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, 2011.

[152] M. Zeppelzauer, "Automated Detection of Elephants in Wildlife Video," *EURASIP Journal on Image and Video Processing*, vol. 46, no. 1, pp. 1–44, 2013.

[153] C. Christoudias, B. Georgescu, and P. Meer, "Synergism in Low Level Vision," in *International Conference on Pattern Recognition (ICPR)*, (Quebec, Canada), pp. 150–155, 2002.

[154] T. Serre, H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, and A. D. Steele, "Automated Home-Cage Behavioral Phenotyping of Mice," tech. rep., Department of Brain and Cognitive Sciences, McGovern Institue, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2009.

[155] H. Jhuang, T. Serre, L. Wolf, T. Poggio, and C. Science, "A Biologically Inspired System for Action Recognition," in *International Conference on Computer Vision (ICCV)*, (Rio de Janeiro, Brazil), 2007.

[156] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," in *Computer Vision and Pattern Recognition (CVPR)*, (Fort Collins, CO, USA), 1999.

[157] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *DARPA Image Understanding Workshop*, (Vancouver, Canada), pp. 168–175, 1981.

[158] E. Simoncelli and D. Heeger, "A Model of Neural Responses in Visual Area MT," *Vision Research*, vol. 38, no. 5, pp. 743–761, 1998.

[159] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-Based Particle Filter for Tracking Multiple Interacting Targets," in *European Conference on Computer Vision (ECCV)*, (Prague, Czech Republic), pp. 1–12, 2004.

[160] S. Oh, J. Rehg, T. Balch, and F. Dellaert, "Learning and Inference in Parametric Switching Linear Dynamic Systems," in *International Conference on Computer Vision (ICCV)*, (Beijing, China), pp. 1161–1168, Ieee, 2005.

[161] A. Veeraraghavan, S. Member, R. Chellappa, and M. Srinivasan, "Shape-and-Behavior-Encoded Tracking of Bee Dances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 463–476, 2008.

[162] V. Pavlovic and J. Rehg, J. M.and MacCormick, "Learning Switching Linear Models of Human Motion," in *Advances in Neural Information Processing Systems*, (Denver, CO, USA), pp. 981–987, 2000.

[163] T. Khan, Z. Balch and F. Dellaert, "A Rao-Blackwellized Particle Filter for EigenTracking," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Washington, DC, USA), pp. II–980 – II–986 Vol.2, 2004.

[164] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 2008.

[165] A. Feldman and T. Balch, "Automatic Identification of Bee Movement Using Human Trainable Models of Behavior," in *International Workshop on the Mathematics and Algorithms of Social Insects*, (Atlanta, GA, USA), pp. 53–60, 2003.

[166] D. P. Gibson, N. W. Campbell, and B. T. Thomas, "Quadruped Gait Analysis Using Sparse Motion Information," in *International Conference on Image Processing (ICIP)*, (Barcelona, Spain), 2003.

[167] D. Tweed and A. Calway, "Tracking Many Objects Using Subordinated Condensation," in *British Machine Vision Conference (BMVC)*, (Cardiff, UK), pp. 26.1–26.10, 2002.

[168] D. Tweed and A. Calway, "Tracking Multiple Animals in Wildlife Footage," in *International Conference on Pattern Recognition (ICPR)*, vol. 2, (Quebec City, Canada), pp. 24–27, 2002.

[169] M. Hindell, M. Lea, and C. Hull, "The Effects of Flipper Bands on Adult Survival Rate and Reproduction in the Royal Penguin Eudyptes Schlegeli," *IBIS*, vol. 138, pp. 557–560, 1996.

[170] I. Cuthill, "Field Experiments in Animal Behaviour, Methods and Ethics," *Animal Behaviour*, vol. 42, pp. 1007–1014, 1991.

[171] M. Daly, M. Wilson, P. Behrends, and L. Jacobs, "Sexually Differentiated Effects of Radio Transmitters on Predation Risk and Behavior in Kangaroo Rats Dipodomys Merriami," *Canadian Journal of Zoology*, vol. 70, pp. 1851–1855, 1992.

[172] J. Peterson, "An Identification System for Zebra (Equus Burchelli Gray)," *East African Wildlife Journal*, vol. 10, pp. 59–63, 1972.

[173] J. Foster, "The Giraffe of Nairobi National Park: Home Range, Sex Ratios, The Herd, and Food," *East African Wildlife Journal*, vol. 4, pp. 139–148, 1966.

[174] G. Schaller, *The Serengeti Lion: A Study of Predator-Prey Relations*. Chicago: University of Chicago Press, 1972.

[175] D. Caldwell, "Evidence of Home Range of an Atlantic Bottlenose Dolphin," *J. Mammal*, vol. 36, no. 2, pp. 304–305, 1955.

[176] A. Irvine and R. Wells, "Results of Attempts to Tag Atlantic Bottlenose Dolphins, Tursiops Truncatus," *Cetology*, vol. 13, pp. 1–5, 1972.

[177] B. Würsig and M. Würsig, "The Photographic Determination of Group Size, Composition, and Stability of Coastal Porpoises (Tursiops Truncatus)," *Science*, vol. 198, pp. 755–756, 1977.

[178] B. Würsig, "Occurrence and Group Organisation of Atlantic Bottlenose Porpoises (Tursiops Truncatus) in an Argentine Bay," *The Biological Bulletin*, vol. 154, pp. 348–359, 1978.

[179] D. T. Bolger, T. A. Morrison, B. Vance, D. Lee, and H. Farid, "A Computer-Assisted System for Photographic Mark-Recapture Analysis," *Methods in Ecology and Evolution*, vol. 3, pp. 813–822, Oct. 2012.

[180] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan, "HotSpotter - Patterned Species Instance Recognition," in *IEEE Workshop an Applications of Computer Vision (WACV)*, (Clearwater Beach, Florida, USA), pp. 230–237, 2013.

[181] L. Hiby and P. Lovell, "Computer Aided Matching of Natural Markings: A Prototype System for Grey Seals," *Report of the International Whaling Commission*, vol. 1, no. 12, pp. 57–61, 1990.

[182] L. Hiby, P. Lovell, N. Patil, N. S. Kumar, A. M. Gopalaswamy, and K. U. Karanth, "A Tiger Cannot Change Its Stripes: Using a Three-Dimensional Model to Match Images of Living Tigers and Tiger Skins.," *Biology Letters*, vol. 5, pp. 383–386, June 2009.

[183] C. Town, A. Marshall, and N. Sethasathien, "Manta Matcher: Automated Photographic Identification of Manta Rays using Keypoint Features," *Ecology and Evolution*, vol. 3, pp. 1902–1914, July 2013.

[184] M. Lahiri, R. Warungu, D. I. Rubenstein, T. Y. Berger-Wolf, and C. Tantipathananandh, "Biometric Animal Databases from Field Photographs: Identification of Individual Zebra in the Wild," in *ACM International Conference on Multimedia Retrieval (ICMR)*, (Trento, Italy), 2011.

[185] J. Duyck, C. Finn, A. Hutcheon, P. Vera, J. Salas, and S. Ravela, "Sloop: A Pattern Retrieval Engine for Individual Animal Identification," *Pattern Recognition*, vol. 2015, no. 48, pp. 1059–1073, 2015.

[186] A. Noviyanto and A. M. Arymurthy, "Automatic Cattle Identification based on Muzzle Photo Using Speed-Up Robust Features Approach," in *European Conference on Computer Science (ECCS)*, (Paris, France), pp. 110–114, 2012.

[187] A. I. Awad, A. E. Hassanien, and H. M. Zawbaa, "A Cattle Identification Approach Using Live Captured Muzzle Print Images," *Advances in Security of Information and Communication Networks*, vol. 381, pp. 143–152, 2013.

[188] A. I. Awad, H. M. Zawbaa, H. A. Mahmoud, E. H. H. A. Nabi, R. H. Fayed, and A. E. Hassanien, "A Robust Cattle Identification Scheme Using Muzzle Print Images," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, (Krakow, Poland), pp. 529–534, 2013.

[189] C. Musgrave and J. L. Cambier, "System and Method of Animal Identification and Animal Transactions Authorization using Iris Patterns," 2002.

[190] U. Gonzales-Barron, G. Corkery, B. Barry, F. Butler, K. McDonnell, and S. Ward, "Assessment of Retinal Recognition Technology as a Biometric Method for Sheep Identification," *Computers and Electronics in Agriculture*, vol. 60, pp. 156–166, Mar. 2008.

[191] U. Gonzales-Barron, F. Butler, K. McDonnell, and S. Ward, "The End of the Identity Crisis? Andvances in Biometric Markers for Animal Identification," *Irish Veterinary Journal*, vol. 62, no. 3, pp. 204–208, 2009.

[192] B. Würsig and T. Jefferson, "Methods of Photo-Identification for Small Cetaceans," *In Individual Recognition of Cetaceans: Use of Photo-Identification and Other Techniques to Estimate Population Parameters, Reports of the International Whaling Commission (Special Issue)*, vol. 12, no. 5, pp. 43–52, 1990.

[193] R. Defran, G. Shultz, and D. Weller, "A Technique for the Photographic Identification and Cataloging of Dorsal Fins of the Bottlenose Dolphin (Tursiops truncatus)," tech. rep., Reports of the International Whaling Commission (Special Issue 12), 1990.

[194] D. Weller, *Global and Regional Variation in the Biology and Behavior of Bottlenose Dolphins*. Phd thesis, Department of Wildlife and Fisheries Sciences, Texas A&M University, 1998.

[195] G. R. Hillman, H. Tagare, K. Elder, A. Drobyshevski, D. Weller, B. Wtirsig, and M. M. Project, "Shape Descriptors Computed from Photographs of Dolphin Dorsal Fins for Use as Database Indices," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, vol. 20, (Hong Kong), pp. 970–973, 1998.

[196] K. Arbter, W. Snyder, H. Burkhardt, and G. Hirzinger, "Application of Affine-Invariant Fourier Descriptors to Recognition of 3D Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 12, pp. 640–647, 1998.

[197] A. Kreho, N. Kehtarnavaz, B. Araabi, G. Hillman, B. Würsig, and D. Weller, "Assisting Manual Dolphin Identification by Computer Extraction of Dorsal Ratio," *Annals of Biomedical Engineering*, vol. 27, no. 6, pp. 830–838, 1999.

[198] D. Marr and E. Hildreth, "Theory of Edge Detection," *Proceedings of the Royal Society of London*, vol. 207, pp. 187–217, 1980.

[199] F. Bergholm, "Edge Focusing," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 9, pp. 726–741, 1987.

[200] B. N. Araabi, N. Kehtarnavaz, T. McKinney, G. Hillman, and B. Würsig, "A String Matching Computer-Assisted System for Dolphin Photoidentification," *Annals of Biomedical Engineering*, vol. 28, pp. 1269–1279, Jan. 2000.

[201] C. Gope, N. Kehtarnavaz, G. Hillman, and B. Würsig, "An Affine Invariant Curve Matching Method for Photo-Identification of Marine Mammals," *Pattern Recognition*, vol. 38, pp. 125–132, Jan. 2004.

[202] E. Mortensen and W. Barrett, "Interactive Segmentation with Intelligent Scissors," *Graphical Models and Image Processing*, vol. 60, no. 5, pp. 349–384, 1998.

[203] A. Ardovini, L. Cinque, and E. Sangineto, "Identifying Elephant Photos by Multi-Curve Matching," *Pattern Recognition*, vol. 41, pp. 1867–1877, June 2008.

[204] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 8, no. 6, pp. 679–698, 1986.

[205] A. L. Blackmer, S. K. Anderson, and M. T. Weinrich, "Temporal Variability in Features used to Photo-Identify Humpback Whales," *Marine Mammal Science*, vol. 16, no. 2, pp. 338–354, 2000.

[206] T. Burghardt, "A General Introduction to Visual Animal Biometrics," tech. rep., Visual Information Laboratory, University of Bristol, Bristol, 2012.

[207] M. J. Kelly and C. Biology, "Computer-Aided Photograph Matching in Studies Using Individual Identification: An Example From Serengeti Cheetahs," *Journal of Mammology*, vol. 82, no. 2, pp. 440–449, 2001.

[208] T. A. Morrison, J. Yoshizaki, J. D. Nichols, and D. T. Bolger, "Estimating Survival in Photographic Capture-Recapture Studies: Overcoming Misidentification Error," *Methods in Ecology and Evolution*, vol. 2, pp. 454–463, Oct. 2011.

[209] N. Kehtarnavaz, V. Peddigari, and C. Chandan, "Photo-Identification of Humpback and Gray Whales Using Affine Moment Invariants," *Image Analysis*, vol. 2749, pp. 109–116, 2003.

[210] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[211] J. Flusser and T. Suk, "Pattern Recognition by Affine Moment Invariants," *Pattern Recognition*, vol. 26, pp. 167–174, 1993.

[212] C. Gope, N. Kehtarnavaz, and G. Hillman, "Zernike Moment Invariants Based Photo-Identification using Fisher Discriminant Model," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, vol. 2, (San Francisco, California, USA), pp. 1455–1458, Jan. 2004.

[213] A. Khotanzad and Y. Hong, "Invariant Image Recognition by Zernike Moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 12, no. 5, pp. 489–497, 1990.

[214] E. Ranguelova and M. Huiskes, "Towards computer-assisted photo-identification of humpback whales," in *International Conference on Image Processing (ICIP)*, vol. 1, (Singapore), 2004.

[215] E. Ranguelova and E. Pauwels, "Saliency Detection and Matching Strategy for Photo-Identification of Humpback Whales," in *International Conference on Graphics, Vision, and Image Processing (GVIP)*, no. December, (Cairo, Egypt), pp. 81–88, 2005.

[216] H. Krijger, *Individual Zebra Identification*. PhD thesis, Rhodes University, South Africa, 2002.

[217] H. Krijger, G. Foster, and S. Bangay, "Designing a Framework for Animal Identification," tech. rep., Computer Science Department, Rhodes University, 2002.

[218] M. Sonka, H. Vaclav, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. Cengage Learning Emea, 2nd edition ed., 1998.

[219] A. Marzal and E. Vidal, "Computation of Normalized Edit Distance and Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 15, pp. 926–932, 1993.

[220] S. Ravela and L. R. Gamble, "On Recognizing Individual Salamanders," in *Asian Conference on Computer Vision (ACCV)*, (Jeju Island, South Korea), pp. 742—-747, 2004.

[221] L. Gamble, S. Ravela, and K. McGarigal, "Multi-Scale Features for Identifying Individuals in Large Biological Databases: An Application of Pattern Recognition Technology to the Marbled Salamander Ambystoma Opacum," *Journal of Applied Ecology*, vol. 45, pp. 170–180, Aug. 2008.

[222] R. Rao and D. Ballard, "Object Indexing Using an Iconic Sparse Distributed Memory," in *International Conference on Computer Vision*, (Boston, MA, USA), pp. 24–31, 1995.

[223] B. M. Har Romeny, *Geometry Driven Diffusion in Computer Vision, Kluwer Academic Publishers*. Springer, 1994.

[224] B. Schiele and J. Crowley, "Object Recognition Using Multidimensional Receptive Field Histograms," in *European Conference on Computer Vision (ECCV)*, (Cambridge, U.K.), 1996.

[225] S. Ravela, J. Duyck, and C. Finn, "Vision-Based Biometrics for Conservation," *Lecture Notes in Computer Science: Pattern Recognition*, vol. 7914, pp. 10–19, 2013.

[226] S. Hoque, F. Deravi, and D. Arts, "ZOOMETRICS a Biometric Identification of Wildlife Using Natural Body Marks," *International Journal of Bio-Science and Bio-Technology*, vol. 3, no. 3, pp. 45–54, 2011.

[227] M. A. H. B. Azhar, S. Hoque, and F. Deravi, "Automatic Identification of Wildlife using Local Binary Patterns," in *IET Conference on Image Processing (IPR)*, (London,UK), pp. 1 – 6, Iet, 2012.

[228] F. Cannavo, G. Nunnari, I. Kale, and F. B. Tek, "Texture Recognition for Frog Identification," in *ACM International Workshop on Multimedia Analysis for Ecological Data (MAED)*, no. 1, (Nara, Japan), pp. 25–29, 2012.

[229] G. Matheron, *Randoms Sets and Integral Equation.* John Wiley & Sons, 1978.

[230] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *International Conference on Computer Vision (ICCV)*, (Barcelona, Spain), pp. 2564–2571, 2011.

[231] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *International Conference on Computer Vision (ICCV)*, (Barcelona, Spain), pp. 2548–2555, 2011.

[232] M. Calonder, V. Lepetit, C. Strecha, and P. Fual, "Brief: Binary Robust Independent Elementary Features," in *European Conference on Computer Vision (ECCV)*, Crete, Greece, pp. 778–792, 2010.

[233] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast Retina Keypoint," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Providence, RI, USA), pp. 510–517, 2012.

[234] E. Pauwels, P. de Zeeuw, and D. Bounantony, "Leatherbacks Matching by Automated Image Recognition," *Advances in Data Mining, Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*, vol. 5077, pp. 417–425, 2008.

[235] M. Fischler and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[236] N. F. Bendik, T. a. Morrison, A. G. Gluesenkamp, M. S. Sanders, and L. J. O'Donnell, "Computer-Assisted Photo Identification Outperforms Visible Implant Elastomers in an Endangered Salamander, Eurycea Tonkawae," *PloS one*, vol. 8, p. e59424, Jan. 2013.

[237] T. Morrison and D. Bolger, "Wet Season Range Fidelity in a Tropical Migratory Ungulate," *Journal of Animal Ecology*, vol. 81, no. 3, pp. 543–552, 2012.

[238] C. Silpa-Anan and R. Hartley, "Optimised KD-Trees for Fast Image Descriptor Matching," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Anchorage, AK, USA), pp. 1–8, 2008.

[239] S. McCann and D. Lowe, "Local Naive Bayes Nearest Neighbor for Image Classification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Colorado Springs, CO, USA), pp. 3650–3656, 2011.

[240] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," *Graphic Gems IV*, vol. San Diego: Academic Press Professional, pp. 474–485, 1994.

[241] Z. Arzoumanian, J. Holmberg, and B. Norman, "An Astronomical Pattern-Matching Algorithm for Computer-Aided Identification of Whale Sharks Rhincodon Typus," *Journal of Applied Ecology*, vol. 42, pp. 999–1011, Dec. 2005.

[242] E. Groth, "A Pattern-Matching Algorithm for Two-Dimensional Coordinate Lists," *Astronomical Journal*, vol. 91, pp. 1244–1248, 1986.

[243] A. van Tienhoven, J. D. Hartog, R. Reijns, and V. Peddemors, "A Computer-Aided Program for Pattern-matching of Natural Marks on the Spotted Raggedtooth Shark," *Journal of Applied Ecology*, vol. 44, pp. 273–280, 2007.

[244] J. Holmberg, B. Norman, and Z. Arzoumanian, "Estimating Population Size, Structure, and Residency Time for Whale Sharks Rhincodon Typus Through Collaborative Photo-Identification," *Endangered Species Research*, vol. 7, pp. 39–53, Apr. 2009.

[245] C. W. Speed, M. G. Meekan, and C. J. Bradshaw, "Spot the Match - Wildlife Photo-Identification Using Information Theory," *Frontiers in Zoology*, vol. 4, no. 2, pp. 1–11, 2007.

[246] C. J. R. Anderson, N. D. V. Lobo, J. D. Roth, and J. M. Waterman, "Computer-Aided Photo-Identification System with an Application to Polar Bears Based on Whisker Spot Patterns," *Journal of Mammology*, vol. 91, no. 6, pp. 1350–1359, 2010.

[247] G. Borgefors, "Distance Transformations in Digital Images," *Computer Vision, Graphics, and Image Processing*, vol. 34, pp. 344–371, 1986.

[248] A. B. Albu, G. Wiebe, P. Govindarajulu, C. Engelstoft, and K. Ovaska, "Towards Automatic Model-Based Identification of Individual Sharp-Tailed Snakes from Natural Body Markings," in *International Conference on Pattern Recognition (ICPR)*, (Tamper, Florida, USA), 2008.

[249] T. Burghardt and N. Campbell, "Individual Animal Identification Using Visual Biometrics on Deformable Coat Patterns," in *International Conference on Computer Vision Systems (ICVS)*, (Bielefeld, Germany), 2007.

[250] S. Belongie, J. Malik, and J. Puzicha, "Shape Context: A New Descriptor for Shape Matching and Object Recognition," *Neural Information Processing Systems (NIPS)*, pp. 831–837, 2000.

[251] H. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistic Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[252] R. Sherley, T. Burghardt, P. Barham, N. Campbell, and I. Cuthill, "Spotting the Difference: Towards Fully-Automated Population Monitoring of African Penguins Spheniscus Demersus," *Endangered Species Research*, vol. 11, pp. 101–111, Mar. 2010.

[253] C. Vincent, L. Meynier, and V. Ridoux, "Photoidentification in Grey Seals: Legibility and Stability of Natural Markings," *Mammalia*, vol. 65, no. 3, pp. 363–372, 2001.

[254] H. Waye, "Can a Tiger Change Its Spots? A Test of The Stability of Spot Patterns for Identification of Individual Tiger Salamanders (Ambystoma Tigrinum)," *Herpetological Conservation and Biology*, vol. 8, no. 2, pp. 419–425, 2013.

[255] M. Domeier and N. Nasby-Lucas, "Annual Re-Sightings of Photographically Identified White Sharks (Carcharodon Carcharias) at an Eastern Pacific Aggregation Site (Guadalupe Island, Mexico)," *Marine Biology*, vol. 150, pp. 977–984, 2006.

[256] C. L. Dudgeon, M. J. Noad, and J. M. Lanyon, "Abundance and Demography of a Seasonal Aggregation of Zebra Sharks Stegostoma Fasciatum," *Marine Ecology Progress Series*, vol. 368, pp. 269–281, 2008.

[257] H. Pratt and J. Carrier, "A Review of Elasmobranch Reproductive Behaviour with a Case Study on the Nurse Shark, Ginglymostoma Cirratum," *Environmental Biology of Fishes*, vol. 60, pp. 157–188, 2001.

[258] A. Castro and R. Rosa, "Use of Natural Marks on Population Estimates of the Nurse Shark, Ginglymostoma Cirratum, at Atol das Rocas Biological Reserve, Brazil," *Environmental Biology of Fishes*, vol. 72, pp. 213–221, 2005.

[259] A. Marshall and M. Bennett, "The Frequency and Effect of Shark-Inflicted Bite Injuries to the Reef Manta Ray (Manta Alfredi)," *African Journal of Marine Science*, vol. 32, pp. 573–580, 2010.

[260] K. P. K. Reddy and R. Aravind, "Measurement of Asymmetry of Stripe Patterns in Animals," in *International Conference on Signal Processing and Communications (SPCOM)*, (Bangalore, India), pp. 1–5, Ieee, July 2012.

[261] H. T. Kim, Y. Ikeda, and H. L. Choi, "The Identification of Japanese Black Cattle by Their Faces," *Asian-Australasian Journal of Animal Sciences*, vol. 18, no. 6, pp. 868–872, 2005.

[262] G. Corkery, U. Gonzales-Barron, F. Butler, K. M. Donnell, and S. Ward, "A Preliminary Investigation on Face Recognition as a Biometric Identifier of Sheep," *American Society of Agricultural and Biological Engineers (ASABE)*, vol. 50, no. 1, pp. 1–8, 2007.

[263] M. S. Bartlett, J. Movellan, and T. Sejnowski, "Face Recognition by Independent Component Analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1463, 2002.

[264] J. L. A. Salle, Q. Wheeler, P. Jackway, S. Winterton, and D. Lovell, "Accelerating Taxonomic Discovery Through Automated Character Extraction," *Zootaxa*, vol. 2217, pp. 43–55, 2009.

[265] K. J. Gaston and M. A. O'Neill, "Automated Species Identification: Why Not?," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 359, pp. 655–67, Apr. 2004.

[266] V. Steinhage, "Automated Identification of Bee Species in Biodiversity Information Systems," *Computer Science for Environmental Protection*, vol. 1, no. 339-344, pp. 1–6, 2000.

[267] K. N. Russell, M. T. Do, J. C. Huff, and N. I. Platnick, "Introducing SPIDA-Web: Wavelets, Neural Networks and Internet Accessibility in an Image-Based Automated Identification System," in *Automated Taxon Identification in Systemtatics: Theory, Approaches, and Applications* (N. E. MacLeod, ed.), ch. Chapter 9, pp. 131–152, 2007.

[268] M. O'Neill, "DAISY: A Practical Tool for Semi-Automated Species Identification," tech. rep., Department of Biology, Newcastle, UK, Newcastle, UK, 2010.

[269] N. MacLeod, "Automated Taxon Identification in Systematics: Theory, Approaches and Applications," *Systematics Association Special Volume*, vol. 74, 2007.

[270] T. Arbuckle, S. Schröder, V. Steinhage, and D. Wittmann, "Biodiversity Informatics in Action: Identification and Monitoring of Bee Species using ABIS," in *International Symposium on Informatics for Environmental Protection (EnviroInfo)*, (Zurich, Switzerland), pp. 425–430, 2001.

[271] S. Schröder, *Automatisierte Identifikation von Bienenarten (Apidae, Hymenoptera) anhand ihres Flügelgeäders durch Methoden der digitalen Bildverarbeitung und der statistischen Klassifikation.* Phd thesis, University of Bonn, 2001.

[272] M. Do, J. Harp, and K. Norris, "A Test of a Pattern Recognition System for Identification of Spiders," *Bulletin of Entomological Research*, vol. 89, pp. 217–224, 1999.

[273] O'Neill, *DAISY: A Practical Computer-Based Tool for Semi-Automated Species Identification, Automated Taxon Identification in Systematics, Theory Approaches and Applications*, ch. 7, pp. 101–114. N. MacLeod (Ed.) CRC Press, 2007.

[274] M. O'Neill, I. Gauld, K. Gaston, and P. Weeks, "DAISY: An Automated Invertebrate Identification System Using Holistic Vision Techniques," in *Inaugural Meeting BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)*, pp. 13–22, 2000.

[275] A. Tofilski, "Using Geometric Morphometrics and Standard Morphometry to Discriminate Three Honeybee Subspecies," *Apidologie*, vol. 39, no. 5, pp. 558–563, 2008.

[276] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual Categorization with Bags of Keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[277] D. A. Lytle, G. Marinez-Munoz, W. Zhang, N. Larios, L. Shapiro, R. Paasch, A. Moldenke, E. N. Mortensen, S. Todorivic, and T. G. Dietterich, "Automated Processing and Identification of Benthic Invertebrate Samples," *Journal of the North American Benthological Society*, vol. 29, no. 3, pp. 867–874, 2010.

[278] L. Breiman, "Randomforest," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[279] A. Utasi, "Local Appearance Feature Based Classification of the Theraphosidae Family," in *Visual Observation and Analysis of Animal and Insect Behavior (VAIB)*, (Tsukuba, Japan), 2012.

[280] G. Burghouts and J. Geusebroek, "Performance Evaluation of Local Colour Invariants," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 48–62, 2009.

[281] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[282] M. Rodrigues, "Automatic Fish Species Classification Based on Robust Feature Extraction Techniques and Artificial Immune Systems," in *International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, no. i, (Liverpool, UK), pp. 1518–1525, 2010.

[283] L. de Castro and F. von Zuben, "aiNet: An Artificial Immune Network for Data Analysis," *Data Mining: A Heuristic Approach*, pp. 231–259, 2001.

[284] G. Bezerra, T. Barra, L. Castro, and F. Zuben, "Adaptive Radius Immune Algorithm for Data Clustering," *Artificial Immune Systems, LNCS*, vol. 3627, pp. 290–303, 2005.

[285] C. Spampinato, D. Giordano, R. Di Salvo, Y.-H. Chen-Burger, R. B. Fisher, and G. Nadarajan, "Automatic Fish Classification for Underwater Species Behavior Understanding," in *ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, (Firenze, Italy), pp. 45–50, ACM Press, 2010.

[286] C. Spampinato, Y. Chen-Burger, G. Nadarajan, and R. Fisher, "Detecting, Tracking and Counting Fish in Low Quality Unconstrained Underwater Videos," in *VISAPP*, pp. 514–519, 2008.

[287] S. Abbasi, F. Mokhtarian, and J. Kittler, "Curvature Scale Space Image in Shape Similarity Retrieval," *Multimedia Systems*, vol. 7, no. 6, pp. 467–476, 1999.

[288] M. Kouda, M. Morimoto, and K. Fujii, "A Face Identification Method of Non-Native Animals for Intelligent Trap," in *Conference on Machine Vision Applications (MVA)*, no. 4, (Nara, Japan), pp. 426–429, 2011.

[289] M. J. Wilber, W. J. Scheirer, P. Leitne, BrianBoult, J. Zott, D. Reinke, D. Delaney, and T. Bolt, "Animal Recognition in the Mojave Desert: Vision Tools for Field Biologists," in *Workshop on the Applications of Computer Vision (WACV)*, (Clearwater Beach, Florida, USA), 2013.

[290] H. M. Afkham, A. T. Targhi, J.-O. Eklundh, and A. Pronobis, "Joint Visual Vocabulary for Animal Classification," in *International Conference on Pattern Recognition (ICPR)*, (Tamper, Florida, USA), pp. 1–4, Dec. 2008.

[291] M. Varma and A. Zisserman, "Texture Classification: Are Filter Banks Necessary?," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[292] X. Yu, J. Wang, R. Kays, P. a. Jansen, T. Wang, and T. Huang, "Automated Identification of Animal Species in Camera Trap Images," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 52, pp. 1–10, 2013.

[293] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear Spatial Pyramidmatching Using Sparse Coding for Image Classification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Miami, Florida, USA), pp. 1794–1801, 2009.

[294] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-Constrained Linear Coding for Image Classification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA, USA), pp. 3360–3367, 2010.

[295] S. Gong, S. J. McKenna, and A. Psarrou, *Dynamic Vision: From Images to Face Recognition*. London, UK: Imperial College Press, 2000.

[296] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[297] M. Abidi, J. Ahlberg, B. Amberg, H. H. Bülthoff, S. Baker, R. Basri, W. Bian, V. Blanz, M. Brauckmann, C. Busch, H. Chang, R. Chellappa, D. Chu, J. F. Cohn, T. Cootes, D. W. Cunningham, X. Ding, M. Du, E. Efraty, R. Gross, P. Grother, B. Guo, A. Hadid, T. Huang, D. Jacobs, A. K. Jain, I. Kakadiaris, T. Kanade, R. Knothe, A. Koschan, J.-K. Kämäräinen, S. Z. Li, X. Liu, Z. Liu, J. B. Martinkauppi, I. Matthwes, R. Micheals, B. Moghaddam, A. J. O'Toole, I. S. Pandzic, S. Pankanti, U. Park, G. Passalis, and P. Perakis, *Handbook of Face Recognition*. London: Springer London Limited 2011, second edi ed., 2011.

[298] R. Chellappa, C. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proceedings of the IEEE*, vol. 83, pp. 705–740, 1995.

[299] W. Zhao, R. Chellappa, and P. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[300] R. Jafri and H. R. Arabnia, "A Survey of Face Recognition Techniques," *Journal of Information Processing Systems*, vol. 5, no. 2, pp. 41–68, 2009.

[301] A. S. Tolba, A. H. El-Baz, and A. El-Harby, "Face Recognition: A Literature Review," *International Journal of Information and Communication Engineering*, vol. 2, no. 2, pp. 88–103, 2006.

[302] V. Vijayakumari, "Face Recognition Techniques: A Survey," *World Journal of Computer Application and Technology*, vol. 1, no. 2, pp. 41–50, 2013.

[303] J. Steffens, E. Elagin, and H. Neven, "PersonSpotter - Fast and Robust System for Human Detection, Tracking and Recognition," in *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, (Washington, DC, USA), 1998.

[304] T. Choudhry, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal Person Recognition Using Unconstrained Audio and Video," in *International Conference on Audio and Video-Based Person Authentication*, 1999.

[305] J. Sivic, M. Everingham, and A. Zisserman, "Who are you? - Learning Person Specific Classifiers from Video," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1145–1152, 2009.

[306] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My Name Is...Buffy - Automatic Naming of Characters in TV Video," in *British Machine Vision Conference (BMVC)*, 2006.

[307] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based Face Recognition on Real-World Data," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, Ieee, Oct. 2007.

[308] H. Ekenel, J. Stallkamp, and R. Stiefelhagen, "A Video-Based Door Monitoring System Using Local Appearance-Based Face Models," *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 596–608, 2010.

[309] T. Kanade, *Computer Recognition of Human Faces.* Basel, Switzerland, and Stuttgart, Germany: Birkhauser, 1973.

[310] M. D. Kelly, "Visual Identification of People by Computer," tech. rep., Stanford AI Project, Stanford, CA, 1970.

[311] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71 – 86, 1991.

[312] E. J. Candes and M. B. Wakin, "An Introduction To Compressive Sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, Mar. 2008.

[313] M. Yang, L. Zhang, and J. Yang, "Robust sparse coding for face recognition," in *Computer and Pattern Recognition (CVPR)*, no. 1, pp. 625–632, Ieee, June 2011.

[314] D. Yang, S. Chen, Y. Chen, and Y. Yan, "Using Head Patch Pattern as a Reliable Biometric Character for Noninvasive Individual Recognition of an Endangered Pitviper Protobothrops Mangshanensis," *Asian Herpetological Research*, vol. 4, no. 2, pp. 134–139, 2013.

[315] H. K. Ekenel and R. Stiefelhagen, "Local Appearance Based Face Recognition Using Discrete Cosine Transform," in *European Signal Processing Conference (EUSIPCO)*, 2005.

[316] H. K. Ekenel, *A Robust Face Recognition Algorithm for Real-World Applications*. PhD thesis, Universitaet Fridericiana zu Karlsruhe (TH), 2009.

[317] T. Ahonen and A. Hadid, "Face recognition with local binary patterns," in *European Conference on Computer Vision (ECCV)*, pp. 469–481, 2004.

[318] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 786–791, Ieee, 2005.

[319] I. J. Cox, J. Ghosn, and P. N. Yianilos, "Feature-Based Face Recognition Using Mixture-Distance," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA, USA), 1996.

[320] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, 1993.

[321] G. Sukthankar, "Face Recognition: A Critical Look at Biologically-Inspired Approaches," tech. rep., Carnegie Mellon University, Pittsburgh, PA, USA, 2000.

[322] L. Harmon, M. Khan, R. Lasch, and P. Raming, "Machine Identification of Human Faces," *Pattern Recognition*, vol. 13, pp. 97–110, 1981.

[323] S. Harmon, L.D. Kuo, P. Raming, and U. Raudkivi, "Identification of Human Face Profiles by Computers," *Pattern Recognition*, vol. 10, pp. 301–312, 1978.

[324] G. Kaufman and K. Breeding, "Automatic Recognition of Human Faces from Profile Silhouettes," *IEEE Transactions On Systems Man And Cybernetics (SMC)*, vol. 6, pp. 113–121, 1976.

[325] Z. Liposcak and S. Loncaric, "A Scale-Space Approach to Face Recognition from Profiles," in *8th International Conference on Computer Analysis of Images and Patterns*, 1999.

[326] Z. Liposcak and S. Loncaric, "Face Recognition from Profiles using Morphological Signature Transform," in *21st International Conference Information Technology Interfaces*, (Pula, Croatia), 1999.

[327] V. Bruce, M. Burtona, and N. Dencha, "What's distinctive about a distinctive face?," *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, vol. 47, no. 1, pp. 119–141, 1994.

[328] V. Bruce, P. Hancock, and A. Burton, "Human Face Perception and Identification," in *Face Recognition* (H. Wechsler, P. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, eds.), vol. 163 of *NATO ASI Series*, pp. 51–72, Springer Berlin Heidelberg, 1998.

[329] J. Sergent, "Microgenesis of Face Perception," in *Aspects of Face Processing* (H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, eds.), vol. 28 of *NATO ASI Series*, pp. 17–33, Springer Netherlands, 1986.

[330] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994.

[331] P. Penev and J. Atick, "Local Feature Analysis: A General Statistical Theory for Objecct Representation," *Network: Computation in Neural Systems*, vol. 7, no. 3, pp. 477–500, 1996.

[332] A. Lanitis, T. C.J., and T. Cootes, "Automatic Face Identification System using Flexible Appearance Models," *Image and Vision Computing*, vol. 13, pp. 393–401, 1995.

[333] G. Edwards, C. Taylor, and T. Cootes, "Learning to Identify and Track Faces in Image Sequences," in *International Conference on AutomaticFace and Gesture Recognition*, 1998.

[334] J. Huang, B. Heisele, and V. Blanz, "Component-Based Face Recognition with 3D Morphable Models," in *International Conference on Audio- and Video-Based Person Authentication*, 2003.

[335] F. B. t. Haar and R. Veltkamp, *3D Morphable Models for Face Surface Analysis and Recognition*, pp. 119–147. John Wiley & Sons SingaporePte Ltd, 2013.

[336] I. Masi, G. Lisanti, A. D. Bagdanov, P. Pala, and A. Del Bimbo, "Using 3D Models to Recognize 2D Faces in the Wild," in *CVPR International Workshop on Socially Intelligent Surveillance and Monitoring (SISM)*, (Portland, OR, USA), 2013.

[337] S. Lin, S. Kung, and L. Lin, "Face Recognition/Detection by Probabilistic Decision-Based Neural Network," *IEEE Transactions on Neural Networks*, vol. 8, pp. 114–132, 1997.

[338] M. Er, S. Wu, J. Lu, and L.-T. Hock, "Face Recognition with Radial Basis Function (RBF) Neural Networks," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 697–710, 2002.

[339] G. Hinton, "Learning Multiple Layers of Representation," *Trends in Cognitive Sciences*, vol. 11, pp. 428–434, 2007.

[340] G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[341] G. E. Hinton, "Deep Belief Networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[342] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[343] C. Küblbeck and A. Ernst, "Face Detection and Tracking in Video Sequences Using the Modified Census Transformation," *Image and Vision Computing*, vol. 24, no. 6, pp. 564–572, 2006.

[344] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, The Edinburgh Building, Cambridge, UK, 2nd ed., 2004.

[345] W. Liao and T. Young, "Texture Classification Using Uniform Extended Local Ternary Patterns," in *IEEE International Symposium on Multimedia (ISM)*, pp. 191–195, 2010.

[346] P. Yang, S. Shan, W. Gao, S. Z. Li, and D. Zhang, "Face Recogniton Using Ada-Boosted Gabor Features," in *IEEE International Conference on Automatic Face and Gesture Recogniton*, (Korea), 2004.

[347] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, New Jersey, USA: Princeton University Press, 1991.

[348] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is Nearest Neighbor Meaningful?," in *International Conference on Database Theory (ICDT)*, pp. 217–235, 1999.

[349] Y. Moses, Y. Adini, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," in *European Conference on Computer Vision (ECCV)*, pp. 286–296, 1994.

[350] N. J. Nilsson, "Introduction to Machine Learning," tech. rep., Department of Computer Science, Stanford University, 2005.

[351] S. Wang, N. Zhang, M. Sun, and C. Zhou, "The Analysis of Parameters t and k of LPP in Several Famous Face Databases," *Lecture Notes in Computer Science*, vol. 6729, pp. 333–339, 2011.

[352] A. Yang, S. Sastry, A. Ganesh, and Y. Ma, "Fast l1-Minimization Algorithms and an Application in Robust Face Recognition: A Review," in *IEEE International Conference on Image Processing (ICIP)*, 2010.

[353] M. Figueiredo, R. Nowak, and S. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[354] D. Donoho, M. Elad, and V. Temlyakov, "Stable Recovery of Sparse Overcomplete RRepresentation on the Presence of Noise," *IEEE Transactions on Information Theory*, vol. 52, pp. 6–18, 2006.

[355] J. Tropp, "Greed is Good: Algorithmic Results for Sparse Approximation," *IEEE Transactions on Information Theory*, vol. 50, pp. 2231–2242, 2004.

[356] Y. Gao, Y. Wang, X. Feng, and X. Zhou, "Face Recognition using Most Discriminative Local and Global Features," in *International Conference on Pattern Recognition (ICPR)*, pp. 351–354, Ieee, 2006.

[357] S. Knerr, L. Personnaz, and G. Dreyfus, *Neurocomputing: Algorithms, Architectures and Applications.*, ch. Single-Layer Learning Revisited: A Stepwise Procedure for Building and Training a Neural Network. Springer-Verlag, 1990.

[358] T.-F. Wu, C.-J. Lin, and R. Weng, "Probability Estimates for Multi-Class Classification by Pairwise Coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[359] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[360] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrica*, vol. 34, no. 4, pp. 321–377, 1936.

[361] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods.," *Neural computation*, vol. 16, pp. 2639–64, Dec. 2004.

[362] J. Zhao, Y. Fan, and W. Fan, "Fusion of Global and Local Feature Using KCCA for Automatic Target Recognition," in *Fifth International Conference on Image and Graphics*, pp. 958–962, Ieee, Sept. 2009.

[363] H. Boström, "Feature vs. Classifier Fusion for Predictive Data Mining a Case Study in Pesticide Classification," in *International Conference on Information Fusion (FUSION)*, (Québec, Canada), 2007.

[364] C. Lip and D. Ramli, "Comparative Study on Feature, Score and Decision Level Fusion Schemes for Robust Multibiometric Systems," in *Frontiers in Computer Education* (S. Sambath and E. Zhu, eds.), vol. 133 of *Advances in Intelligent and Soft Computing*, pp. 941–948, Springer Berlin Heidelberg, 2012.

[365] D. Ruta and B. Gabrys, "An Overview of Classifier Fusion Methods," *Computing and Information Systems*, vol. 7, pp. 1–10, 2000.

[366] B. Gökberk, A. A. Salah, and L. Akarun, "Rank-based decision fusion for 3D shape-based face recognition," in *International Conference on Audio-and Video-Based Biometric Person Identification (AVBPA)*, pp. 1019–1028, 2005.

[367] E. Murphy-Chutorian and M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[368] W. Morgenstern, "Keyframe-Selektion zur Erhöhung der Erkennungsrate von Primaten in Videos," Master's thesis, Technical University of Ilmenau, Institute of Media Technology, 2014.

[369] R. Liu, Z. Li, and J. Jia, "Image Partial Blur Detection and Classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[370] P. Bex and W. Makous, "Spatial Frequency, Phase, and the Contrast of Natural Images," *Journal of the Optical Society of America A*, vol. 19, no. 6, pp. 1096–1106, 2002.

[371] D. Field, " Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells," *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[372] D. Field and N. Brady, "Visual Sensitivity, Blur and the Sources of Variability in the Amplitude Spectra of Natural Scenes," *Vision Research*, vol. 37, no. 23, pp. 3367–3383, 1997.

[373] O. Hilliges, P. Kunath, A. Pryakhin, A. Butz, and H. Kriegel, "Browsing and Sorting Digital Pictures Using Automatic Image Classification and Quality Analysis," *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, vol. 4552, pp. 882–891, 2007.

[374] M. Graef, *Entwicklung und Implementierung von Modulen zur Erkennung technischer Bildfehler*. Bachelor Thesis, Hochschule für Technik, Wirtschaft und Kultur Leipzig, 2011.

[375] D. P. Curtin, *The Textbook of Digital Photography*. Marblehead, Massachusetts, USA: Shortcourses and Photocourse Publishing Programs, 2007.

[376] P. Mahalanobis, "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[377] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications*. Imperial College Press, 2000.

[378] S. Eddy, "Maximum Likelihood Fitting of Extreme Value Distributions," tech. rep., Department of Genetics, Washington School of Medicine, 1997.

[379] T. Ypma, "Historical Development of the Newton-Raphson Method," *Society for Industrial and Applied Mathematics (SIAM) Review*, vol. 37, no. 4, pp. 531–551, 1995.

[380] K. Atkinson, *An Introduction to Numerical Analysis*. John Wiley and Sons, Inc., New York, 1989.

[381] A. Ernst, T. Ruf, and C. Küblbeck, "A Modular Framework to Detect and Analyze Faces for Audience Measurement Systems," in *2nd Workshop on Pervasive Advertising 2009 in conjunction with Informatik 2009*, (Lübeck, Germany), pp. 3941–3953, 2009.

[382] A. Jain, R. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[383] K. Lee, J. Ho, and D. D. Kriegman, "Acquiring Linear Subspaces for Face Recognition under Variable Lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[384] P. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET Database and Evaluation Procedure for Face Recognition Algorithms," *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295–306, 1998.

[385] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 25, no. 9, pp. 1063–1074, 2003.

[386] N. Dahm and Y. Gao, "A Novel Pose Invariant Face Recongition Approach Using a 2D-3D Searching Strategy," in *International Conference on Pattern Recognition (ICPR)*, 2010.

[387] A. Rama, F. Tarres, D. Onofrio, and S. Tubaro, "Mixed 2D-3D Information for Pose Estimation and Face Recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[388] X. Zhang, Y. Gao, and M. Leung, "Recognizing Rotated Faces From Frontal and Side Views: An Approach Towards Effective Use of Mugshot Databases," *Information Forensics and Security*, vol. 3, no. 4, pp. 684–697, 2008.

[389] F. Hajati, A. A. Raie, and Y. Gao, "Pose-Invariant 2.5 D Face Recognition using Geodesic Texture Warping," in *International Conference on Control Automation Robotics and Vision (ICARCV)*, 2010.

[390] D. G. Kendall, "A Survey of the Statistical Theory of Shape," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[391] F. L. Bookstein, *Morphometric Tools for Landmark Data.* Cambridge, UK: Cambridge University Press, 1991.

[392] J. Foley, A. van Dam, S. Feiner, and J. Hughes, *Computer Graphics: Principles and Practice*, ch. 16: Illmuniation and Shading, pp. 721–814. Addison-Wesley Publishing Company, Inc., 2006.

[393] X. Zhang and Y. Gao, "Heterogeneous Specular and Diffuse 3-D Surface Approximation for Face Recognition Across Pose," *Information Forensics and Security*, vol. 7, no. 2, pp. 506–517, 2012.

[394] P. J. Phillips, S. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1104, 2000.

[395] P. J. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone, "Face Recognition Vendor Test 2002: Evaluation Report," nistir 6965, National Institute of Standards and Technology, 2003.

[396] P. J. Phillips, P. Grother, and R. Micheals, *Handbook of Face Recognition*, ch. Chapter 21: Evaluation Methods in Face Recognition, pp. 551–574. Springer-Verlag London Limited, 2011.

[397] B. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 5814, pp. 972–976, 2007.

[398] C. Zhu, F. Wen, and J. Sun, "A Rank-Order Distance Based Clustering Algorithm for Face Tagging," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[399] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 3 ed., 2006.

[400] S. Pizer, E. Amburn, J. Austing, R. Cromartie, A. Geselowitz, T. Greer, B. Romeny, J. Zimmerman, and K. Zuiderveld, "Adaptive Histogram Equalization and its Variations," *Computer Vision, Graphics and Image Processing*, vol. 39, pp. 355–368, 1987.

[401] D. Jobson, Z. Rahman, and G. Woodell, "A Multiscale Retinex for Bridging the Gap Between Color Images and the Human Observations of Scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[402] E. Land and J. McCann, "Lightness and Retinex Theory," *Journal of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[403] V. Štruc and N. Pavešić, *Photometric Normalization Techniques for Illumination Invariance*, pp. 279–300. IGI-Global, 2011.

[404] E. Tola, V. Lepetit, and P. Fua, "Daisy: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, 2010.

[405] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a Local Binary Descriptor Very Fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012.

[406] E. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM (Association for Computing Machinery)*, vol. 13, no. 6, pp. 377–387, 1970.

[407] D. Chamberlin and R. Boyce, "SEQUEL: A Structured English Query Language," in *ACM SIGFIDET Workshop on Data Description, Access, and Control (Association for Computing Machinery)*, (New York, NY, USA), pp. 249–264, 1974.

[408] A.-S. Crunchant, M. Egerer, K. Zuberbuhler, K. Corogones, V. Leinert, L. Kulik, and H. Kuehl, "Man or Machine: Application and Potential of Biometric Software to Generate Baseline Occurrence Estimates in Chimpanzee Communities," *American Journal of Primatology*, 2015 (submitted, not yet published).

[409] S. Rana, W. Liu, M. Lazarescu, and S. Venkatesh, "Efficient tensor based face recognition," in *19th International Conference on Pattern Recognition*, pp. 1–4, Ieee, Dec. 2008.

[410] S. Rana, W. Liu, M. Lazarescu, and S. Venkatesh, "A unified tensor framework for face recognition," *Pattern Recognition*, vol. 42, pp. 2850–2862, Nov. 2009.

[411] G. Hamerley and C. Elkan, "Learning the K in K-Means," in *Neural Information Processing Systems*, 2003.

[412] F. Böhme, "Automatische Gruppierung ähnlicher Gesichter unbekannter Individuen," Master's thesis, TU Ilmenau, 2013.

[413] K. Aizawa, S. Baker, T. E. Boult, S. Chaudhuri, M.-C. Chiang, U. B. Desai, N. P. Galatsanos, T. Kanade, A. Katsaggelos, N. Kaulgud, T. Komatsu, J. Mateos, R. J. Micheals, R. Molina, S. Peleg, and D. Rajan, *Super-Resolution Imaging*. Kluwer Academic Publishers, 2002.

[414] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and Challenges in Super-Resolution," *International Journal of Imaging Systems and Technology*, vol. 14, no. 2, pp. 47–57, 2004.

[415] C. Liu and D. Sun, "On Bayesian Adaptive Video Super Resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 2, pp. 346–360, 2013.

[416] B. Gunturk and A. Batur, "Eigenface-Domain Super-Resolution for Face Recognition," *IEEE Transactions on Image Processing*, vol. 12, no. 5, pp. 597–606, 2003.

[417] F. Wheeler, X. Liu, and P. Tu, "Multi-Frame Super-Resolution for Face Recognition," in *IEEE International Conference on Biometrics: Theory and Applications*, pp. 1–6, 2007.

[418] P. Hennings-Yeomans, S. Baker, and B. Vijaya Kumar, "Recognition of Low-Resolution Faces Using Multiple Still Images and Multiple Cameras," in *IEEE International Conference on Biometrics: Theory, Systems, and Applications*, IEEE Computer Society, September 2008.

[419] S. Baker and T. Kanade, "Hallucinating Faces," in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 83–88, 2000.

[420] J. Yang, H. Tang, Y. Ma, and T. Huan, "Face Hallucination via Sparse Coding," in *IEEE International Conference on Image Processing*, pp. 1264–1267, 2008.

[421] B. B. Mandelbrot, *The Fractal Geometry of Nature*. W.H. Freeman Company, New York, N.Y, 1977.

[422] B. Lashermes, S. Jaffard, and P. Abry, "Wavelet Leader Based Multifractal Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2010, Jan. 2005.

[423] H. Wendt, S. G. Roux, S. Jaffard, and P. Abry, "Wavelet Leaders and Bootstrap for Multifractal Analysis of Images," *Signal Processing*, vol. 89, pp. 1100–1114, June 2009.

[424] Y. Xu, X. Yang, and H. Ling, "A New Texture Descriptor using Multifractal Analysis in Multi-Orientation Wavelet Pyramid," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, no. 60603022, 2010.

[425] J. L. Dickinson, B. Zuckerberg, and D. N. Bonter, "Citizen Science as an Ecological Research Tool: Challenges and Benefits," *Annual Review of Ecology , Evolution, and Systematics*, vol. 41, pp. 149–172, 2010.

[426] "citizen n. and adj", *Oxford University Press.* OED online, March 2015 (accessed April 07, 2015).

[427] J. A. Ahumada, C. E. F. Silva, K. Gajapersad, C. Hallam, J. Hurtado, E. Martin, A. McWilliam, B. Mugerwa, T. O'Brien, F. Rovero, D. Sheil, W. R. Spironello, N. Winarni, and S. J. Andelman, "Community Structure and Diversity of Tropical Forest Mammals: Data From a Global Camera Trap Network," *Philosophical Transactions of the Royal Society of London B*, vol. 366, no. 1578, pp. 2703–2711, 2011.

[428] B. Hoppe-Dominik, H. Kühl, G. Radl, and F. Fischer, "Long-Term Monitoring of Large Rainforest Mammals in the Biosphere Reserve of Taï National Park, Côte d'Ivoire," *African Journal of Ecology*, vol. 49, no. 4, pp. 450–458, 2011.

Eidesstattliche Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Weiterhin habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalte der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer Prüfungsbehörde vorgelegt.

Ich bin darauf hingewiesen worden, dass die Unrichtigkeit der vorstehenden Erklärung als Täuschungsversuch angesehen wird und den erfolglosen Abbruch des Promotionsverfahrens zur Folge hat.

Ilmenau, den 05.05.2015 Alexander Loos