Technische Universität München

Bildverstehen und Intelligente Autonome Systeme

Facial Expression Recognition With A Three-Dimensional Face Model

Christoph Mayer

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Johann Schlichter
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Bernd Radig (i.R.)
	2. Univ.-Prof. Gudrun J. Klinker, Ph.D.

Die Dissertation wurde am 04.07.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 02.12.2011 angenommen.

Abstract

Over the last decades, speed and data storage capabilities of computers have steadily increased, allowing to process ever more data in less time and to fulfill ever more complex tasks. However, the tools users rely on to specify the task have hardly changed during that period. Traditionally, keyboard, mouse, screen and speakers are used for communication between user and machine. Unfortunately, those tools are frequently considered unintuitive and uncomfortable, especially by technically inexperienced persons. Therefore, recent research considers more advanced human-machine communication, often inspired by well-known human-human communication techniques.

In this thesis, facial expressions, especially those conveying basic human emotions, are investigated to improve human-machine interaction, since they are one of the most important communication modalities for humans. Automated facial expression recognition systems face a number of characteristic challenges. Firstly, in contrast to artificial objects, which have a well-defined structure, human faces differ a lot with respect to appearance and shape. Secondly, obtaining natural training data is difficult, especially for facial configurations expressing emotions like sadness or fear. Therefore, publicly available databases consist of acted facial expressions and are biased by the author's design decisions. Finally, evaluating trained algorithms towards real-world behavior is challenging, again due to the artificial conditions of available image data.

We tackle each of these challenges separately: We propose a novel image preprocessing procedure that highlights facial components like face skin, eyebrows and lips, rendering it specifically suitable for face image analysis tasks. Furthermore, we propose a novel face model fitting strategy that is applied to a three-dimensional face model, which is highly suitable for facial expression recognition. Our fitting strategy is evaluated on images that have been collected from the media and therefore have not been taken with a computer vision in mind. Finally, rather than training our classifier from only a single database, we use several databases in our evaluation. We present a novel evaluation strategy that trains classifiers on one database and tests them on another to ward against overspecialization. We present applications of our technique in the area of cognitive technical systems. We embed facial expression recognition a vivid dialog with a robot head and demonstrate the advantage compared to a non-emotional dialog via a user study.

Zusammenfassung

Verarbeitungsgeschwindigkeit und Speicherkapazität von Rechnern sind über die letzten Jahrzehnte stetig gewachsen und gestatten die Verarbeitung von immer mehr Daten in immer kürzerer Zeit um immer komplexere Aufgaben zu erfüllen. Auf der anderen Seite haben sich die Hilfsmittel, welche Benutzern zur Verfügung stehen, um die Aufgaben zu spezifizieren, in derselben Zeit kaum verändert. Traditionell werden Tastatur, Maus, Bildschirm und Lautsprecher für die Kommunikation zwischen Mensch und Maschine benutzt. Leider werden diese Hilfsmittel, gerade von technisch unerfahrenen Benutzern, oft als unintuitiv und unkomfortabel wahrgenommen. Daher untersucht neuere Forschung höher entwickelte Methoden zur Mensch-Maschine Kommunikation, die oft von klassisch menschlicher Kommunikation inspiriert sind.

In dieser Arbeit werden Mimiken, speziell jene, welche Information über grundlegende Emotionen vermitteln, untersucht, um die Mensch-Maschine Kommunikation zu verbessern. Systeme zur automatisierten Erkennung menschlicher Mimik werden mit einer Anzahl von charakteristischen Herausforderungen konfrontiert. Erstens variieren menschliche Gesichter, im Gegensatz zu künstlich hergestellten Objekten, die eine definierte Struktur haben, in Bezug auf Aussehen und Form. Zweitens ist es schwierig, natürliche Trainingsdaten zu erhalten, besonders für Mimiken die Trauer oder Furcht widerspiegeln. Daher besteht zugängliches Trainingsmaterial aus geschauspielerten Mimiken und ist durch Designentscheidungen der Datenbankauthoren vorbelastet. Zuletzt ist die Evaluation von gelernten Algorithmen hinsichtlich Verhalten in der wirklichen Welt wegen der Künstlichkeit dieser Daten schwierig.

Diese Herausforderungen werden separat in Angriff genommen: Eine neue Bildvorverarbeitung hebt Gesichtskomponenten wie die Gesichtshaut, Augenbrauen und Lippen hervor. Weiterhin wird eine neue Strategie zur Anpassung eines Gesichtsmodells, welches besonders gut für den Einsatz zur Mimikerkennung geeignet ist, angewendet. Diese Strategie wurde mit Hilfe von Bildern evaluiert, welche aus den Medien gesammelt wurden und deshalb nicht mit der Absicht aufgenommen wurden, für automatisierte Bildinterpretation verwendet zu werden. Weiterhin werden mehrere Datenbanken zum Trainieren des Klassifikators genutzt, anstatt sich nur auf eine einzelne zu verlassen. Die Klassifikatoren werden mittels eines neuen Verfahrens evaluiert, welches darauf beruht, Klassifikatoren auf einer Datenbank zu trainieren und auf einer Anderen auszuwerten um Überspezialisierung zu vermeiden. Anwendungen dieser Techniken werden im Bereich kognitiver technischer Systeme vorgestellt. Mimikerkennung wird in einen Dialog mit einem Roboterkopf eingebunden, um die Vorteile gegenüber einem Dialog ohne Mimiken durch eine Umfrage darzustellen.

Acknowledgments

Although I created this document on my own, I received much support in conducting the research that led to its creation. First of all, I would like to thank my supervisor, Prof. Dr. Bernd Radig, for his mentoring and helpful assistance in my scientific work. I got to know him as a person of great acumen who often provided well-thought advice and who always established a good working atmosphere. He provided me with the opportunity to work on a fascinating topic in a challenging research area.

Furthermore, I would like to thank Dr. Matthias Wimmer and Dr. Freek Stulp for introducing me to this interesting research topic and to scientific work processes in general. They both helped me a lot, especially at the start of my work. I would also like to thank several colleges with whom I spent many hours inside and outside the lab, and who I consider not only as fellow workers but also as friends: Zahid Riaz, with whom I shared an office for almost two years of work. I greatly enjoyed the long discussions on face image analysis. Martin Eggers, who became my office mate later, and who helped me greatly with his English language skills. Stefan Sosnowski, with whom some of the most fascinating work has been developed. Tobias Rehrl, who often inspired me with fruitful discussions. Finally, Jürgen Blume and Alexander Bannat, for their friendship and support. Additionally, I would also like to thank all the other colleges from the CoTeSys Central Robotics Lab and Intelligent Autonomous Systems Group, who are too many to be personally addressed here.

Many thanks are also directed to the persons who helped me with proofreading this thesis, namely Martin Eggers, Tobias Rehrl, Jürgen Blume and Stefan Sosnowski. Finally, I would like to thank my family, my parents Gerhard and Christine, my sister Judith and my brother Philip, for supporting me steadily during my work and my writing.

Contents

Chapter 1

Introduction

The ability to communicate is, without doubt, of utmost importance to humans and animals alike. Psychologist and philosopher Paul Watzlawick defined five axioms of communication, of which the first states that ”One cannot not communicate” [141]. What Watzlawick refers to here, is that humans transmit information passively even when they do not actively communicate. Humans are not aware of all signals that they transmit and do not actively know all rules they apply to interpret these signals. However, communication also plays a major role in linking technical systems. In contrast to human-human communication, machine-machine communication relies on predefined, designed and therefore well-known rules and specifications. Other than the often black-boxed human communication, machine-machine communication is, given appropriate knowledge, completely transparent. Considering the most obvious application of machine-machine communication, the Internet, it can not be denied that both, human-human communication and machine-machine communication, are integrated parts of our daily live.

Metcalfe’s law states that the usefulness of a network can be approximated by n^2 with n being the number of communication nodes [143]. Although measuring usefulness might be a difficult endeavor, its core idea, to measure the number of connections that can be established with a number of network components, is an intuitive approach. One of the conclusions of the law is that by fusing two separate networks into one, the resulting network is more useful than the simple sum. On the other hand, one of the critics on Metcalfe’s law is that one should measure the number of active connection rather than the number of theoretically possible connections, which again directly leads to the question of interfacing between the two networks. The current drawback of machine communication is that, in order to use it, one has to learn its rules. We learn human communication mechanisms intuitively during childhood, but to understand machine communication reading often lengthy manuals is required for a single machine and studies of years to get a more general view.

Therefore, research has started to consider not only improving machine capacity, but also the way of communication between humans and machines. Fusing machine-machine and human-human communication techniques offers many challenges, since humans are not fully aware of all communication they use themselves. Although humans are able to learn machine communication protocols, it is more desirable that machines adapt or ”learn” human communication protocols, since, in the end, machines are invented to serve humans, not the other way around. It is of high importance to humans to pass information about their emotional state, since it provides important context to verbal communication and heavily influences the behavior of both interaction partners. Psychology and Intercultural Communication investigate the influence of subconsciously passed signals in human-human communication and offer a wide variety of examples of how communication that works perfectly fine on the verbal level can be ruined by such non-verbal signals through seemingly small details.

In this thesis, we inspect one of the most important human communication modalities, facial expressions, and demonstrate, how machines can understand or at least recognize them from video images. Facial expression recognition refers to the task of automatically determining the facial expression of a human from camera images, see Figure 1.1 for two example images. Often, facial expression recognition is performed in a three-step approach: face detection, feature extraction from the located face region and classification of the facial expression visible. We follow the same steps as the overview of our approach in Figure 1.2 shows.

Face detection refers to the challenge of determining the face position within the image, usually by a bounding box. We apply the approach of Viola et al. for this task to determine the face bounding box and the position of eyes in the image [136]. This approach is known for its speed and robustness, and publicly available implementations exist that have been trained on a large number of example images. A preprocessing step extracts meaningful low-level features to support subsequent high-level image interpretation. This step highlights important image content, specific facial components, and is especially helpful in real-world applications with cluttered background and difficult lighting conditions. Knowledge about the location of the face components provides a more profound basis than pure pixel information.

For facial expression feature extraction, we fit a face model, which provides

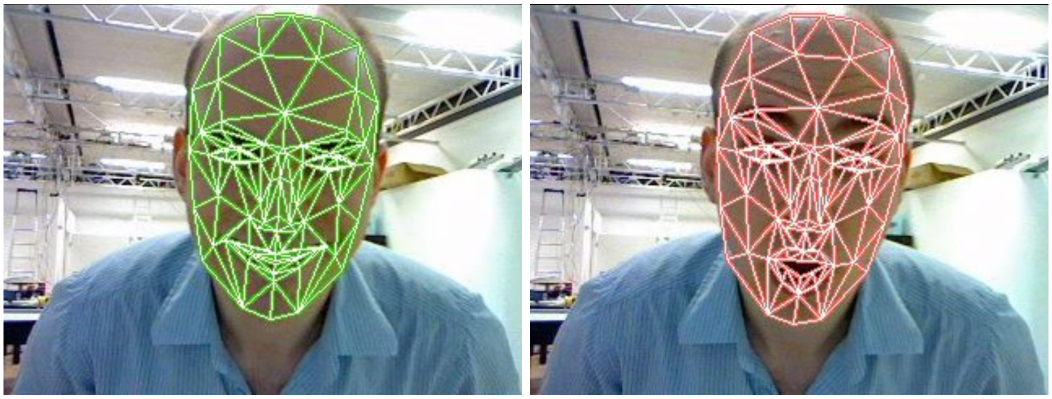


Figure 1.1: We present a system that recognizes facial expressions from camera images.

an abstract representation of valuable information about important face properties, onto the image. The Candide-III face model contains information about the face position in 3D space rather than appearance in the image [2]. Starting with a default parameterization of the model, regression algorithms refine the model parameters iteratively. To obtain representative evaluations, we integrate the ”*Labeled Faces In The Wild*” database, which consists of images that have been taken outside lab or office environments. This database offers face images with a wide variety in head pose, lighting, ethnic background and facial expression.

In the human face, eight muscles are used to generate a large number of facial expressions. However, only a few facial expressions correspond to emotional reactions. Ekman and Friesen found six universal facial expressions that are expressed and interpreted independent of cultural background, age or country of origin all over the world and correspond to a set of emotional states: happiness, surprise, fear, anger, sadness and disgust. The Facial Action Coding System (FACS) precisely describes the muscle activity within a human face that appears during the display of facial expressions. To automatically recognize facial expressions, one either has to infer them directly from the image data, or has to determine the activation of action units in the face and compare them to the list specified by Ekman and Friesen. In both cases, a database of exemplary image data is required to evaluate the algorithm. Today, a number of such databases is available to the community. However, due to the fact that obtaining real image data is difficult, and that humans often find it difficult to act certain facial expressions, the image material is often biased by the expectations or instructions of the authors, rendering an evaluation of how accurate the algorithms perform in out-of-the-lab scenarios difficult. Therefore, we include multiple databases in our evaluation for a fair comparison.

1.1 Contributions

The contributions of this thesis are three-fold and we state them explicitly:

We present a system for facial expression recognition from camera images. We follow the three main steps proposed by Pantic et al. [99] and present algorithms for the most challenging steps: The feature extraction step and the facial expression classification step. Preprocessing of images to highlight important facial components allows for a robust face model fitting. We integrate a face model, which is highly suitable for for facial expression recognition. Classifiers are trained to determine the facial expression from the parameters of the fitted face model.

Furthermore, we present evaluations for implementations of those two steps that are tuned to reflect robustness in real-world scenarios. Machine learning techniques in computer vision require data that reflects the variety of image content. To assure comparability, most approaches for facial expression recognition are evaluated on standard databases. These databases, as far as model fitting is concerned, often restrict the image content with respect to face size, lighting, background or similar context conditions. In contrast, we present evaluations based on images that have been captured in real world conditions without a computer vision application in mind, and that are publicly available for comparison. As far as facial expression recognition is concerned, most approaches are evaluated on a single database. Therefore, algorithms are often tuned towards high recognition rates on this specific database and obtain them at the cost of generalization to other databases. In contrast, we evaluate our algorithms on different databases to obtain evaluations that are not biased by database properties.

Finally, we present practical applications of our algorithms. Although they also contain scenarios for facial expression classification, not all of them directly refer to this topic. Instead, we highlight the role of the proposed approaches to work as a subsystem in a more complex system with a determined functionality, taken from the field of human-machine communication. Furthermore, the requirement for different techniques of evaluation, such as user studies, is demonstrated and an example of such an evaluation is shown.

1.2 Outline

The remainder of this thesis is structured as follows:

- Chapter 2 summarizes challenges, which automated facial expression recognition is confronted with, and that separate it from other computer vision tasks. These challenges are tackled one by one to present solution ideas in several processing steps.
- Chapter 3 outlines related work on facial expression recognition. It provides a categorization of related approaches, identifies open research questions and links the work presented in this thesis to existing research.
- Chapter 4 introduces our novel preprocessing procedure, that creates multi-band images from the raw image data. This image representation is specifically tuned towards face image analysis tasks, because it highlights important facial components like eyebrows or lips.
- Chapter 5 presents our model fitting strategy, that integrates mentioned multi-band images. We train displacement experts that propose parameter updates on single model parameters and evaluate our approach on the ”Labeled Faces in the Wild” database, which offers challenging data.
- Chapter 6 details the determination of facial expressions from model parameters. We include three publicly available databases and train classifiers to recognize seven facial expressions. Our classifiers are evaluated in cross-database evaluation to inspect their generalization capabilities.
- Chapter 7 demonstrates applications of the methods introduced in previous chapters. We establish a human-machine dialog that is guided by facial expression analysis and synthesis. Evaluation is conducted with user studies on human experiment participants.
- Chapter 8 provides a conclusion of this work and hints future work. It reviews and summarizes the conclusions drawn in the other chapters to demonstrate how this work contributes to the state-of-the-art.

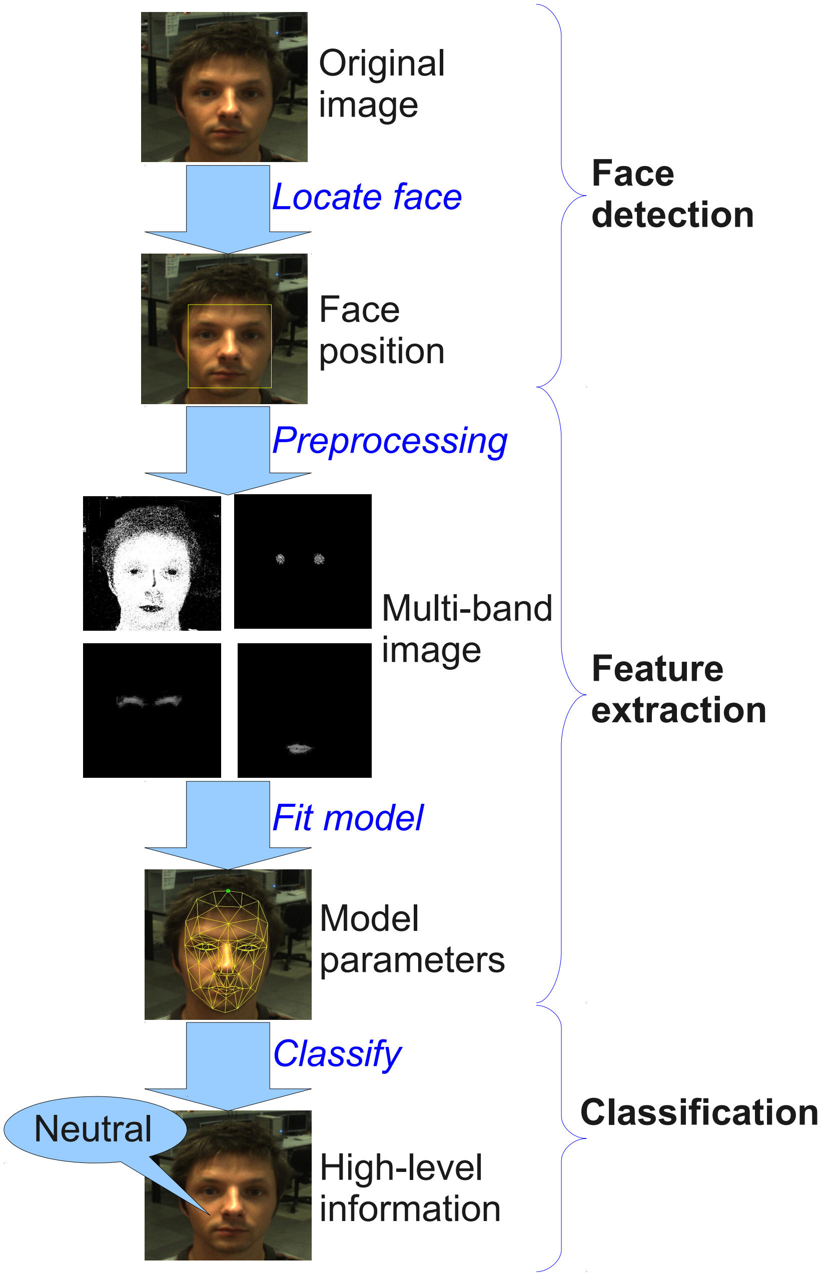


Figure 1.2: Overview of the complete process.

Chapter 2

Problem Statement and Solution Idea

The automated analysis of human faces is confronted with many characteristic challenges compared to other computer vision tasks. Therefore, no other object has received such interest from the community, specifically, since face analysis inherently holds large opportunities for interdisciplinary research with psychologists and neurologists. In this chapter, we will highlight these specific challenges and present approaches that tackle them at different processing steps, starting with the raw image data until the facial expression classification itself.

2.1 Problem Statement

Face image analysis considers the task of automatically obtaining high-level information from images of human faces. The most prominent applications in this area are face identification and facial expression recognition. Face identification determines the identity of the person visible in the image, or determines whether two images refer to the same person. Facial expression recognition estimates the face state according to a semantic interpretation, either regarding the complete face (laughing, crying or talking) or only facial components (closed eyes, opened mouth). There are several characteristics to this area of research that have to be considered.

In contrast to many artificial objects, faces differ in their size, appearance and aspect ratios severely. This holds several benefits for computer vision applications, but at the same time induces several challenges. One of the benefits is that faces are unique which allows to determine a person's identity from their face. Since artificial objects often look alike, other means to distinguish them are required, like product numbers or license plates for cars. Although faces also might look similar, no two faces are exactly the same. We intuitively make use of this fact when using a person's face to identify this person.

On the other hand, from a computer vision point of view, this also requires a certain flexibility and robustness of algorithms for the interpretation of face images. Although some properties of faces are known (they have two eyes, one nose, one mouth etc.), the structure and texture of a specific face has to be extracted online. Often, face models serve as a tool to provide this flexibility while ensuring robustness at the same time. Face models represent the depicted face in a small number of descriptive model parameters and provide an abstraction of the image content. However, model parameters that match the image content have to be calculated in order to provide useful information, a process referred to as ”face model fitting” or ”face alignment”. Further challenges exist due to environment aspects, such as lighting or background. Many applications that involve face image analysis are considered in unstructured or uncontrolled environments. Unfortunately, most publicly available data is not captured in such realistic surroundings, but depicts face images taken in labs and offices, in which the lighting is controllable and the human is always looking straight to the camera. Therefore, training robust algorithms from such data or determining the robustness of an algorithm is not a trivial problem.

Furthermore, facial expressions are used by humans on a very subconscious level. Therefore, specifying them explicitly or providing explicit models for their appearance and context dependency is a difficult endeavor and is often subject to subjective decisions. Research demonstrates, that even humans are capable of classifying short sequences without any context information only at a rate of roughly 75% [146]. Additionally, obtaining natural or real-world data is difficult, especially for facial expressions referring to emotions like sadness or fear. Facial expressions often do not occur pure, but in combinations such as a surprised smile. However, available data mainly consists of acted facial expression, where people are instructed to display facial expression according to database author instructions. Summing up, in contrast to artificial objects, where the appearance and behavior is specified in technical terms during the production and which therefore follow these rules, the recognition and interpretation of facial expressions requires taking several uncertainties into account.

As a consequence, the core challenge when considering face images in computer vision applications is three-fold: Obtaining representative data, creating robust calculation rules and conducting fair evaluations of the algorithms. In the context of face model fitting, these core challenges manifest in the large variety of faces' visual appearance, either due to person-specific aspects, like complexion, hair color or facial hair, or due to facial expressions. In the context of facial expression recognition these core challenges manifest in the variety with which persons depict facial expressions and in the fact that facial expressions are often highly context-related.

2.2 Solution Idea

To compensate for these varieties, in the object of interest as well as in the surrounding, robust calculation rules in the three main processing steps are required. Defining explicit calculation rules in any of these steps manually can be a difficult if not impossible task. Therefore, machine learning is applied in all three steps, allowing to define positive or negative examples. This is often much easier for humans than specifying the calculation rules directly, which is then done by the computer.

We follow a multi-step approach that transforms the raw image data into information about the face visible in the image. Each step is intended to create a more abstract representation of the image, neglecting irrelevant information. In each step, we integrate data for training and testing, that is specifically chosen to reflect the variety of real-world scenarios.

The first step creates multi-band images, that highlight specific facial components. This image representation serves two purposes: It separates the background from the visible face and it provides rough structural information about the face. Since execution time is highly important, we rely on boosted tree stumps.

The second step fits a face model onto the image which reduces the multi-band image data to a number of model parameter values. Parameters that describe face aspects like the opening of eyes or mouth. To fit the model, parameter updates are calculated to the initial parameterization and therefore we face a regression problem rather than a classification task. We integrate image data in this step, that has been captured in a large number of different occasions with various persons, backgrounds and lighting conditions.

In the third step, this information is further reduced and combined to determine the facial expression visible in the image. To compensate for author bias in the training images, we integrate multiple databases and conduct our evaluation cross-database. Cross-database evaluation is an evaluation strategy that obtains the training and test data from different databases, instead of acquiring both as different subset from the same database. We integrate Support Vector Machines for this classification task.

Chapter 3

Related Work

An important step in facial expression research has been taken by Ekman et al. in the 70s, who identified six universal facial expressions that correspond to specific emotions and that are expressed in the same way independent of age, gender or cultural background all over the world [31]. To precisely describe the structure of facial expressions, the same authors published the Facial Action Coding System (FACS), which describes facial expressions by activations of so-called ”action units” in the face. Each action unit refers to a single intra-face movement and to the contraction of specific facial muscles [34]. Much research has been conducted on automatically identifying either the activation of single action units or universal facial expressions.

Systems for face image analysis are divisible in two categories: Systems that determine single facial action (rising a single eyebrow, closing the eyes, opening the mouth, ...) and systems that determine a complete facial expression, usually with a specific semantic meaning (happiness, confusion, sleeping, ...). Approaches in the first category usually propose single facial actions or a set of states for facial components of which some are semantically exclusive and some are independent. For instance, the eyes are either widened or closed, but either state may be combined with an opened mouth. These approaches provide a face state space, yet without any semantic meaning. Approaches in the second category follow a different idea: they provide a description that involves the complete face, usually with a semantic meaning. The semantics mostly refers to a set of mind states, like emotions (happiness, sadness, surprise,...), boredom, stress or even pain. In literature, the detection of facial actions is sometimes also referred to as ”facial expression recognition”. However, we will refer to them as systems for facial action recognition, in order to provide a consistent terminology.

Furthermore, approaches are divided to be either image-based or model based. Image-based approaches rely on the extraction of image features like gradient direction, local binary patterns, Haar-like features or Gabor wavelet responses. Model-based approaches integrate an additional step by fitting a model to the image and inspecting the parameters of the fitted model instead of the image content directly. Models represent a-priori knowledge about the face and allow to form an abstraction of the image content that neglects unimportant image properties.

A very early survey on the topic of face recognition and facial expression recognition is presented by Samal and Iyengar [113]. Their work is focused on human capabilities in this area and presents automated approaches in several subtasks, like face detection, identification and representation. Almost ten years later, Pantic et al. authored one of the most frequently referred to surveys [99]. They also identify some important steps that most available approaches rely on, and already present a selection of systems that tackle all of them. The approaches presented cover model-based approaches and image-based approaches alike. Since, at this time, no database for facial expression analysis with wide acceptance had been published, different approaches are presented but not compared for their accuracy. Only a few years later, Fasel et al. present another survey of this field that very explicitly considers the difference between facial expression recognition and emotion recognition [38]. They still recognize the need for a widely accepted image database but already mention the *Cohn-Kanade Facial Expression Database* (CK database), which will evolve into one of the most utilized databases in this area [63]. Furthermore, they mention that all systems presented still require manual interference by the user for face detection or initialization. This issue is only a minor challenge today, as many approaches for robust face detection have been proposed and this can be considered an (almost) solved problem. A very recent survey is presented by Zeng et al. [155]. They focus on multi-modal affect recognition and present approaches for facial expression recognition, affect recognition from audio or the integration of gestures. They review a large number of approaches on several important, publicly available databases, among them the most often used ones: The Cohn-Kanade Facial Expression Database and the MMI Face Database [63, 96]. Evaluation is usually conducted by sampling training and test data from the same database. Rarely, cross-database evaluation is performed.

3.1 Facial Action Recognition

Mostly, the FACS is used to determine single facial actions by determining the activation of single action units. Manually labeled databases, like the CK database serve as training and test data to evaluate such systems. Their advantage is their flexibility, since combinations of activated action units are indicative for many states of mind, not only emotions but also confusion, fatigue or boredom [155]. However, because manually labeling image sequences is a tedious task, much effort has been taken to automatize this task. One example is presented in our earlier work, which determines the intensity of action unit activations [16, 44].

Another example is presented by De la Torre et al., who aim at assisting professional FACS coders in their work [69]. They take a model-based approach and create person-specific face models, which renders the approach inapplicable to previously unseen data, unless a face model is created for this person. This, however, requires manually labeling image data, again. Since facial actions are inherently dynamic in their movement, they extract temporal segmentation of facial behavior from image data. They demonstrate that their approach works well on image data that has not been taken with a computer vision application in mind and therefore induces challenging context conditions.

An approach that is not person-specific is proposed by Bartlett et al. Their image-based system recognizes the activation of 27 different facial action units by applying Gabor-wavelets to the face region and utilizes SVMs and AdaBoost for classification [78]. They evaluate their system on different databases that have been manually FACS-annotated. One of the databases contains images taken during a real conversation and therefore depicts real instead of acted facial expressions. Participants were asked to convince a neutral person of a specific opinion on a political or social subject. This opinion was either their real opinion or the opposite, and the neutral person had to determine, which was the case.

These systems work on the believe that action unit activations appear independent of each other, or at least do not model the dependencies explicitly. Since this is not a realistic assumption, Tong et al. model the dependencies of action units in a Bayesian Network. They expect that some combinations of action units appear more likely linked than independent [149]. Taking these dependencies into account increases the detection accuracy significantly.

All of the mentioned systems so far consider frontal-view images, since this represents the usual dialog situation and is therefore an intuitive assumption. However, Pantic et al. demonstrate that breaking this assumption might be beneficial to the detection of certain action units and propose a system that uses profile view images for facial action detection [98]. They argue that several facial actions, like pushing the tongue under the lips, are not observable in frontal view images. 15 anatomical landmarks are tracked with a particle filtering method to detect the activation of 27 action units. Most probably inspired by this work, the authors publish the MMI database, which contains frontal-view and profile-view images [96]. In a more recent work, they investigate the temporal dynamics of action unit activations [114, 66]. Boosted Gabor-wavelet features and Hidden Markov Models are combined to determine the neutral, onset, apex and offset stage. They evaluate their approach on the MMI and CK database and also conduct one cross-database evaluation where the MMI database serves as training database and the CK database is used for testing. Observed accuracy are higher on the CK database than across databases.

3.2 Facial Expression Recognition

Much work in recognizing predefined facial expressions rather than single facial actions is dedicated to classifying the six universal facial expression of Ekman et al. [110, 150, 67, 100, 148, 5, 74]. Facial expressions can be considered static, as a fixed face state, or dynamic, as intra-face movement. An example for the first approach is given by Kotsia et al., who determine the facial expression by detecting the activation of several action units and then applying the rules of Ekman et al. [67]. They utilize the CK database and fit a face model to the neutral image in the beginning of each sequence. Then, they track it through the sequences to obtain model parameters for the strongly exaggerated expressions in the end of the sequences. Modified SVMs are trained to determine the activation of certain, selected action units to obtain the facial expression from the rules stated by Ekman and Friesen.

An approach that follows the opposite idea in many aspects is presented by Anderson et al. [4]. Their approach is inspired by the dynamics of facial expressions and they utilize an image-based method to determine face motion from optical flow. Furthermore, they strongly emphasize the real-time capability of their system and also present some example applications like an interactive chat client. However, similar to the approach of Kotsia et al., they train and evaluate, both fully automatized, on the CK database. Inspired by the insight that standard databases are known for its strongly exaggerated facial expressions, Park et. al. aim at determining subtle facial expressions by artificially enhancing the face motion to produce an exaggerated facial expression [100]. They use a face model to extract facial motion vectors for 27 facial feature points and artificially enhance this face motion to produce an exaggerated facial expression. Training and evaluation is conducted on a database that contains four facial expressions (neutral, smile, anger, surprise) of Asian people.

In earlier work, we also presented a similar approach for real-time facial expression recognition in a live demonstration [83]. Evaluation of this system is conducted on the CK database only. A more detailed description and evaluation on two databases, the CK database and the MMI database, is given in [127]. Our current work presented in this thesis extends the approach by presenting an evaluation on three databases and in cross-database evaluation.

Practical application is also a major focus of interest in the work of Whitehill et al. [55]. They focus on one facial expression only and follow an image-based approach to determine whether a person is smiling or not. In previous work, they present a system for smile detection that has been evaluated on two databases, but not in cross-database evaluation [74]. The system relies on AdaBoost and Support Vector Machines to determine the facial expression from convolutions of image data with Gabor energy filters. However, when they test their system in real-world condition, they recognized a large drop in accuracy. Inspired by this observation, they collect a very large database for practical smile detection and train two different classifiers on three different feature sets. They recognize that providing evaluations, which are valid not only on a set of predefined datasets but reflect real-world behavior, is difficult. Recognition of facial expressions on a low resolution, as it might appear in real-world data, is the focus of Shan et al. [121]. They extract local binary patterns from the image data and evaluate their approach on several databases. Mainly, the CK database is used, but results are also reported on the MMI database and meeting recordings. Furthermore, they evaluate their approach across databases, training on the CK database and evaluating on the MMI database and observe a severe loss in accuracy.

Intensive comparison also inspired the work of Sebe et. al. [119], but they inspect the classification technique rather than the training data. In their evaluation, they train 24 different types of classifiers on facial expression data received from a specially designed database to determine which type of classifier is best suited for facial expression classification. They test their results on the CK as well as on a specifically designed Authentic Facial Expression Database.

Estimating the facial expression intensity is considered less frequently in literature, but some research has been conducted. Yang et al. present facial expression intensity estimation with boosted ranking [150]. The core idea of this approach is to determine the intensity by boosted ranking rather than traditional regression techniques. Instead of specifying explicit intensity values for single images, images are only compared for their intensity difference. Therefore, the algorithm is trained to decide for two example images which of the images depicts the stronger facial expression. However, although they report high classification rates, they evaluate their approach only on the CK database, which is known for its exaggerated facial expressions.

Facial expressions that do not directly link to one of the universal emotions have also been considered. Automatic detection of pain has been considered in medical applications [3]. Traditionally, skilled human raters assign labels between 0 (no pain) and 5 (strong pain) to single images. Ashraf et al. use the same rating and obtain human-labeled training data to train SVMs and sequence-based classification.

3.2.1 Conclusion

Almost all of the above mentioned approaches utilize either information extracted from complete image sequences or apex expressions only. While this is very well applicable to database sequences, there are difficulties to be expected if these approaches are applied in real-world conditions, when people express facial expressions in more varying intensity or length. For instance if people are in a relaxed conversation, they might show a smiling face over an extended time, which will not be recognized by sequence-based classification, since no change in the facial expression occurs. Similar problems might occur with approaches that consider apex expressions only. If the system is trained on strongly laughing faces only, it might consider the mentioned smiling face to be still neutral.

Therefore, image-based classification is required that decides for a single image, which facial expression is depicted. Furthermore, the whole variance of intensity of facial expression has to be considered, in order to ensure that the point of transition as it is determined by the classifier matches the intuitive perception of a human. Considering this data is more challenging than focusing on the apex expressions only, but is also a more realistic approach.

In this thesis, we take the next logical step in following these conclusions: Instead of training algorithms that perform well on test images with constrained content, we aim at determining the system performance when these restrictions are released. We specify facial expression labels for single images and consider facial expressions at varying intensities. Furthermore, we conclude that images captured in controlled environments do not reflect the variability of real-world data very well, neither for model fitting nor for facial expression recognition. Therefore, we propose to confront systems with data that helps warding against specializing on single databases instead of rewarding it. Whitehill et al. state that ”It is conceivable that by evaluating performance on these data sets the field of automatic expression recognition could be driving itself into algorithmic ’local maxima’. ” [55]. In this thesis, we propose approaches that aim at avoiding this local maximum, following the roadmap that is depicted in Section 2.2. We propose a novel evaluation strategy that aims at obtaining results that are representative for real-world conditions, instead of achieving ever higher results on well-structured, artificial data.

Chapter 4

Low-level feature extraction

The interpretation of image data is often arranged in multiple steps with every step providing more abstract information to subsequent processing steps. Low-level feature extraction supports the extraction of higher-level information to increase speed and accuracy. Extracted low-level features might include edge intensity [20], Haar-like features [136], optical flow [67], pixel value differences [8] or SIFT features [95, 76] to detect facial expressions [67], fit or track models [95, 20, 67], determine person identity [8] or detect complex objects in image data [95].

In this chapter, we present an image preprocessing procedure, which is specifically tuned to the task of interpreting human face images. Its application supports subsequent image interpretation by segmenting facial components from the rest of the face and the image background. This information is represented in so-called ”multi-band images”. Image-bands can be thought of as additional image channels, but we refrain from this nomenclature in order to avoid confusions. Although most face image interpretation approaches utilize standard image representations in the *red-green-blue* color space or in the *hue-saturation-value* color space, more complex image representations have been proposed, already. We take our nomenclature from Stegmann et al., who compute an image representation, which contains a mixture of pixel values in different color spaces and call this a multi-band image representation [128]. Cootes et al. present an image representation based on image edges, in which the additional image bands reflect edge direction and intensity [20]. Our image bands highlight specific facial components, like lips or eye brows, and provide more semantic information than simple image filtering. Figure 4.1 depicts an example image with its image bands.

The creation of these multi-band images is based on two components: characteristics of human faces and characteristics of a single image. From this information an intermediate step calculates descriptive feature values for single image pixels. They represent the pixel’s spatial location and color value with respect to the pre-estimated spatial distributions and color distributions of various facial components. Therefore, they inherently represent information about the surrounding image and form descriptive information for subsequent classification tasks.

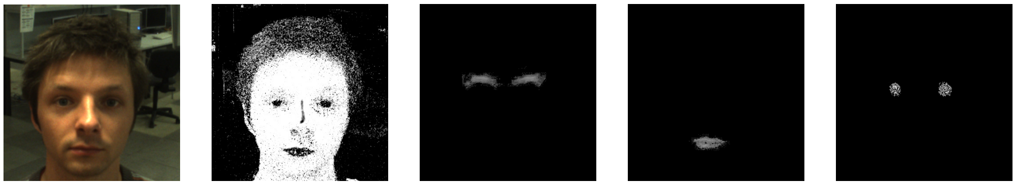


Figure 4.1: Image bands highlight the position of facial components in the image.

We call these pixel features ”adjusted pixel features”, because, in addition to the raw image data, they also take characteristics of the complete image into account. After calculating adjusted pixel features for all image pixels, classifiers are trained on them to decide for a single pixel, which facial components this pixel depicts. Applying these classifiers to all pixels of an image segments facial components from the rest of the face and the image background. Please see Figure 4.2 for an overview of the complete process. The advantage of pixel-based classifiers is that they provide high runtime performance, and, since we provide them with adjusted pixel features, achieve high accuracy at the same time.

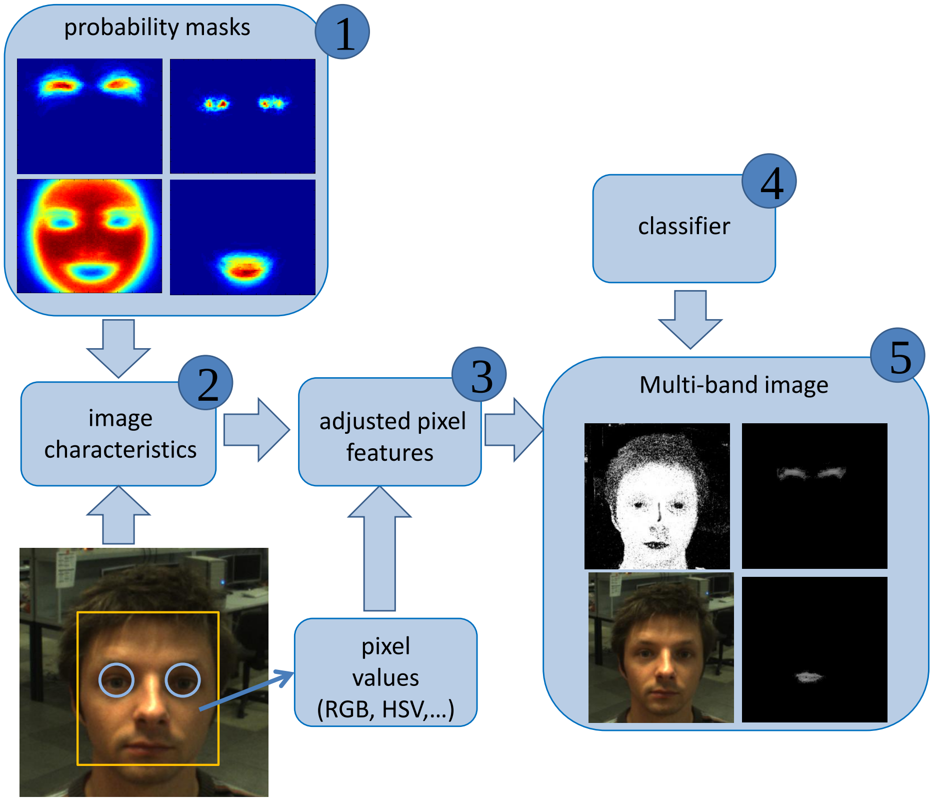


Figure 4.2: The probability masks are determined off-line in the first step. In the second step image characteristics are computed to calculate adjusted pixel features in the third step. Pixel-based classifiers are trained on them in the fourth step. Finally, multi-band image are created in the fifth step.

4.1 Problem Statement

Classifiers for pixel-based segmentation of human faces often rely on we call *static pixel features*. These features consist of information that is extracted from single pixels and raw image data, like the pixel coordinates in the image and the pixel color values. Therefore, they are independent of the surrounding image content. Since only the information of a single pixel is considered in pixel-based approaches, the mapping of the pixel features to the facial component must be robust in order to achieve satisfactory results for the complete image.

Unfortunately, the color of a specific facial component varies significantly throughout a set of random images. The reason is that varying context conditions, such as lighting, camera type and settings, and the person's complexion and ethnic group, make the color of a facial component occupy a large cluster within any color space. Furthermore, color clusters of different facial components may overlap. The pure spatial location of a pixel does not provide any useful information at all, because human faces potentially appear anywhere in the images.

4.2 Solution Idea

In this thesis we propose to calculate so-called *adjusted pixel features* that are adapted to the image content. The basic idea is that, if combined and compared to the surrounding image content, static pixel features still provide useful information. For instance, in a single image all pixels of a certain facial component look similar and share certain color statistics, which eases the task of finding potential pixel candidates for this facial component. Furthermore, the color clusters of different facial components are less likely to overlap, and pixels of the same facial component are located in a small area rather than in the whole image.

Inspired by this insight, we utilize static pixel features to calculate another set of pixel features, which specifically take the image context into consideration. Face properties are represented by a set of probability matrices that provide spatial estimations of facial components. They are calculated inside a square region of interest (ROI), which surrounds the face and is determined by applying a face locater. Then, the entries of the probability masks are computed to reflect the probability that a certain facial components is visible at a specific position in the ROI. We utilize this information to estimate a set of image characteristics, such as an estimation of the color cluster of certain facial components within the given face region. Finally, from the static pixel features and the image characteristics we calculate additional pixel features, the mentioned adjusted pixel features that are adapted to the context conditions. Therefore, they are more suitable to determine facial components, by training classifiers that determine for a single pixel which facial component is depicted.

4.3 Related Work

Often, sophisticated mathematical methods are applied to locate facial components, varying from template matching to model fitting. Although, this chapter rather provides a preprocessing to support such methods, we will still shortly review them as alternatives to find facial components. A straightforward idea is to use templates or shape information [70, 133, 21, 94, 37, 84]. A model of the face or facial component is fit to the image, or component candidates are detected and verified. The models and candidate determination algorithms incorporate the geometric structure of the facial component or the appearance in the image, which guides search for this structure within the face areas, we will see more about that in Chapter 5. These methods tend to show high accuracy but also high execution time. In contrast, our algorithm allows for identifying facial components using simple and quick computation schemes, and rather provides beneficial information for subsequent model fitting. We refer to our earlier work for a summary of our approach [17]. The recent version utilizes AdaBoost instead of SVMs for classification to increase the execution speed.

Classifying skin color is addressed more frequently in the literature than other facial components. Therefore, we review related approaches individually.

4.3.1 Extracting Skin Color

A large region of the human face is covered by skin and therefore, skin color represents an important source of information to various computer vision applications considering human faces. For a recent survey we refer to Phung et al. [102] and Vezhnevets et al. [135]. Vezhnevets et al. categorize the detection techniques as follows:

Nonparametric skin color distribution modeling individually inspects every element of the color space to determine whether it represents skin or not. Rules to make this decision are learned from comprehensive training data, which require a lot of memory, both for learning and for storing the rules. The advantage of these algorithms is that they perform at very high speed. An example of this approach is presented by Jones et al. [59].

Parametric skin color distribution modeling models skin color to be located within a cluster with a specific shape defined by a set of parameters, which limits the required memory to these parameters. The computation of the parametric cluster increases classification time. However, the accuracy decreases because the predefined shape of the cluster does not represent the true color distribution exactly. Common approaches model skin color distribution via a single Gaussian or a mixture of Gaussians [59, 53].

Explicit definition of the skin color cluster uses a set of rules that explicitly define the color cluster without sticking to a predefined geometric shape. Memory requirements are limited by the chosen rules. Most often, this task is accomplished by rule induction algorithms that learn the rules from an annotated training set. Features that are well associated with skin color often allow to obtain accurate rules. An example approach is presented in [102]

Dynamic skin color distribution modeling extends the previously mentioned techniques by additionally considering further image conditions rather than relying on the color of the pixel only. In consequence, the cluster's shape is adapted to the processed image, which improves the skin detection accuracy. The skin color cluster of Soriano et al. looks like the crescent of the moon [126]. They call it *skin locus* and this shape is image-specific.

4.3.2 Identifying Facial Components from Color

There are several applications, in which detailed information about single facial components is required. For instance, lip classifiers provide useful information for speech recognition, speaker authentication and lip tracking [105, 111, 112]. The approach of Leung et al. is very similar to our lip classification approach, estimating clusters that exactly localize the lips in the face [71]. Liew et al. also consider color and spatial features for lip segmentation [73]. However, detecting facial components may also support the detection of the complete face to reduce the error rate. This idea is followed by Hsu et al. [53]. They construct eye and mouth maps to verify each face candidates. In contrast, our approach uses the previously estimated location of the face in order to precisely determine the location of the facial components. A similar approach has recently been published by Beigzahed et al. [84]. They manually define rules to construct the eye and mouth maps for determining mouth and eye candidates. In contrast, our approach applies automatically trained classifiers. Manually annotating images is much more

a) static pixel features (color information)	
$\boldsymbol{x}_c$	color information of the pixel.
b) static pixel features (spatial information)	
$\boldsymbol{x}_s$	The coordinates of the pixel within the image.

Table 4.1: Color and spatial information of the static pixel features.

straightforward and therefore less error-prone than manually constructing the decision rules. Furthermore, including other facial components than eye and mouth, such as iris or eye brows, merely requires annotating the data instead of defining a new set of rules.

4.3.3 Conclusion

All presented approaches rely on color distributions that are determined from a set of example images, but they mostly rely on manually constructed decision rules, in which only the parameters are estimated. For instance, skin color is often expected to cover a small cluster in the HSV color space, or lips are known to be red and separated by an image edge from the surrounding image. The main disadvantage of this approach is that it is difficult to extend to other facial components. An approach that requires only labeling facial components and constructs the decision rules how to locate this component on its own would be preferable, since it is more objective and easier to extend. Furthermore, only a few of the presented approaches adapt the color distribution to the specific image [126]. However, taking the characteristics of a specific image into account, instead of assuming that one set of decision rules works for any image, would greatly increase the robustness. Therefore, we present an approach that refrains from manually constructing decision rules and determines them fully automatically from annotated images. The drawback of this approach is that it requires more annotated data, since not only test data but also training data is needed. However, this is easily compensated for by the gain in robustness.

4.4 Computing the Adjusted Pixel Features

This section details the computation of adjusted pixel features from off-line generated probability masks, image characteristics and static pixel features. Static pixel

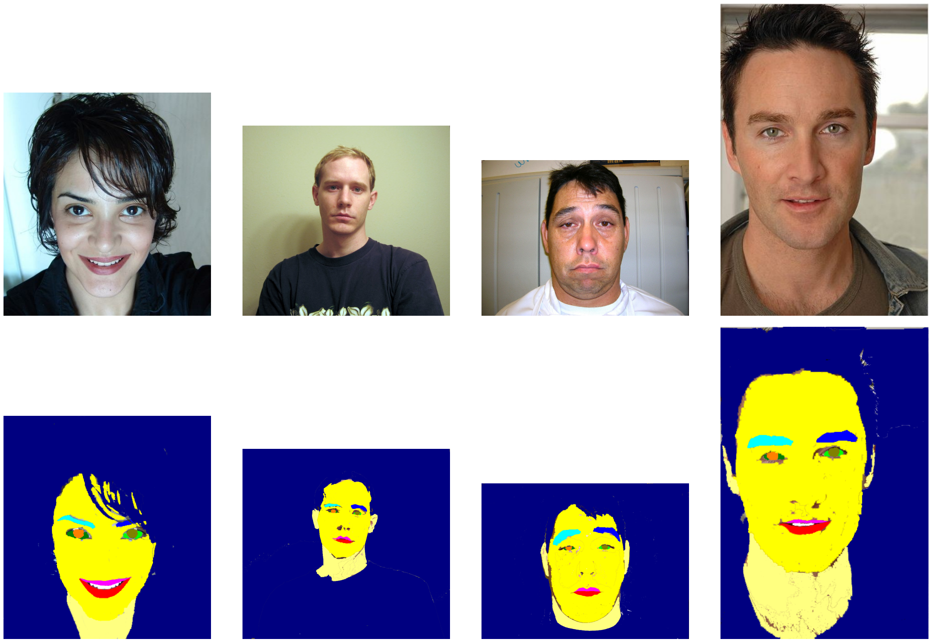


Figure 4.3: Facial components are manually annotated in training images.

features consist of the pixel coordinates and color values of a pixel as summarized in Table 4.1. The static pixel feature, the image characteristics and the adjusted pixel features are categorized to being either spatial or color-related. Color-related features are represented in different color spaces at the same time, such as RGB, NRGB, HSV. However, in order not to overstrain the mathematical notation, the explanation below only handles one color space. Similarly, we consider spatial information in different representations, such as Cartesian coordinates and polar coordinates with different origins.

Both, generating the probability masks and training the pixel-based classifiers, requires annotated training data. Therefore, we collect a large number of images from the internet, depicting person of different age, gender, complexion and background, and manually annotate facial components. Please see Figure 4.3 for some example images and annotations.

4.4.1 Facial Component Masks

To estimate the ROI around the visible face, we apply the face locater of Viola and Jones [136]. We generate a set of matrices from training images, one matrix A^f

per facial component f, that contain the probability for each pixel within the ROI to depict a specific facial component. The matrix values are calculated from the relative frequency of occurrence in the training images. This requires to map every pixel of the mask to the corresponding area of the ROI in all training images. By counting, how often this area depicts the facial component and dividing this number by the total number of training images, we obtain an estimation of the likelihood that the facial component is visible at this location in test images. The idea of these masks is to represent structural properties of the human face: Firstly, they provide a rough estimation of where certain facial components are located within the ROI, for instance that the eyebrows are always located in the upper area or the lips are always located on the lower area. Secondly, they provide information about the relative location of facial components to each other, for instance that the eyebrows are always located above the lips.

After the generation of the masks, we are able to predict the spatial distribution of facial components within the ROI of test images by scaling A^f to the size of the ROI of the given test image. For each pixel $\boldsymbol{x}$, we estimate its probability $b^f_{\boldsymbol{x}}$ to depict the facial component f from the the matrix entry corresponding to the pixel position within the ROI. Figure 4.4 presents probability masks for a set of facial components $f \in \mathcal{F} = \{skin, lips, brows, retina\}$. The generation of the probability masks refers to Step 1 in Figure 4.2.

4.4.2 Image Characteristics

The image characteristics model properties of a single, given image rather than single pixels and therefore have to be calculated only once per image. Since the rough position of the face is an important information, the parameters of the ROI determined by the face locater are directly considered in the image characteristics. Furthermore, we determine the position of the eyes within the ROI to provide information more robust against turned faces, please see Figure 4.5 for a visualization in an example image. Their position within the image as well as their distance also characterize the image content.

Further image characteristics are obtained by exploiting the probability masks. Each computation considers pixel color information $\boldsymbol{x}_c$ or pixel coordinates $\boldsymbol{x}_s$, a pixel's probability $b^f_{\boldsymbol{x}}$ to be part of the facial component f, and the number of pixel $|ROI|$ within the ROI. We model the spatial distribution of every facial component to be Gaussian and compute the distribution parameters $(\boldsymbol{m}^f, S^f)$ as in Equation 4.1.

The color distribution of facial components is also modeled to be Gaussian

a) image characteristics (spatial information)	
$\boldsymbol{m}^f$, S^f	The predicted spatial distribution of the facial component f within the image. We assume this distribution to be Gaussian. Its computation is illustrated in Equation 4.1.
l	The landmark index from the set of landmarks $L = \{NE, SE, NW, NE, C, E\}$ provided by the face locater.

b) image characteristics (color information)	
$\boldsymbol{\mu}^f$, Σ^f	The predicted color distribution of the facial component f within the entire image. We assume this distribution to be Gaussian. Its computation is illustrated in Equation 4.2.

Table 4.2: Color and spatial information of the image characteristics.

with the distribution parameters $(\boldsymbol{\mu}^f, \Sigma^f)$. The pixel values of all pixels that are predicted to be covered by a single facial component according to the probability masks are considered for this computation. As with the spatial distributions, the color distribution parameters $(\boldsymbol{\mu}^f, \Sigma^f)$ calculated from Equation 4.2 of all facial Components contribute to the image characteristics. Table 4.2 summarizes all extracted image characteristics. The calculation of the image characteristics corresponds to Step 2 in Figure 4.2.

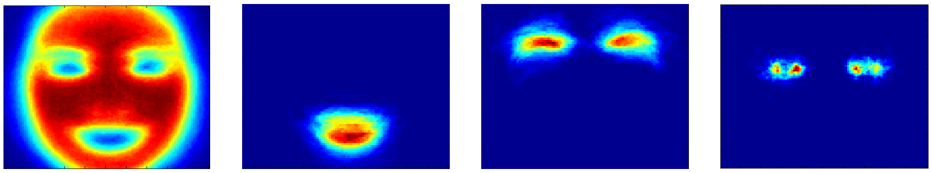


Figure 4.4: We train a set of probability masks that reflect the frequency of occurrence of a specific facial component within the region of interest. In this example, the probability masks for skin, lips, eyebrows and retinas are depicted.

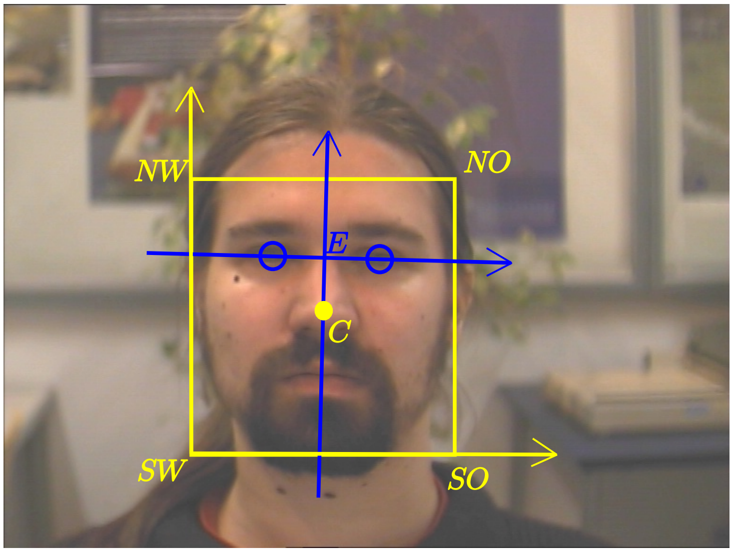


Figure 4.5: Several landmarks in the image serve as reference points for the spatial image characteristics.

$$\begin{aligned} \boldsymbol{m}^f &= \frac{1}{|ROI|} \sum_{\boldsymbol{x} \in ROI} \boldsymbol{x}_s \, b^f_{\boldsymbol{x}}. \\ S^f &= \frac{1}{|ROI|} \sum_{\boldsymbol{x} \in ROI} (\boldsymbol{m}^f - \boldsymbol{x}_s)(\boldsymbol{m}^f - \boldsymbol{x}_s)^T (b^f_{\boldsymbol{x}})^2. \end{aligned} \tag{4.1}$$

$$\begin{aligned} \boldsymbol{\mu}^f &= \frac{1}{|ROI|} \sum_{\boldsymbol{x} \in ROI} \boldsymbol{x}_c \, b^f_{\boldsymbol{x}}. \\ \Sigma^f &= \frac{1}{|ROI|} \sum_{\boldsymbol{x} \in ROI} (\boldsymbol{\mu}^f - \boldsymbol{x}_c)(\boldsymbol{\mu}^f - \boldsymbol{x}_c)^T (b^f_{\boldsymbol{x}})^2. \end{aligned} \tag{4.2}$$

4.4.3 Adjusted Pixel Features

This section presents the calculation of adjusted pixel features from static pixel features and image characteristics. As presented in Equation 4.3, the adjusted spatial pixel features i_l contain the pixel coordinates relative to the image's facial landmark positions $x_l \in L = \{NE, SE, NW, NE, C, E\}$ defined by the ROI. The distance is calculated as the Euclidean distance, normalized by the ROI's side size, as well as the angle difference from polar coordinates. Two reference coordinate systems are used, one defined by the ROI and one defined by the eye positions, see Figure 4.5. The distance measurement is normalized by the interocular distance (the distance between the eyes). Furthermore, adjusted spatial pixel features k^f and ℓ^f represent the location of the pixel relative to the spatial distribution of the facial component f. Again, the Euclidean k^f and the Mahalanobis ℓ^f distance between the pixel's location $\boldsymbol{x}_s$ and the spatial mean $\boldsymbol{m}^f$ of the facial component f are is calculated as given in Equation 4.4 and in Equation 4.5.

$$i_l = |\boldsymbol{x}_s - \boldsymbol{x}_l|. \tag{4.3}$$

$$k^f = |\boldsymbol{x}_s - \boldsymbol{m}^f|. \tag{4.4}$$

$$\ell^f = \sqrt{(\boldsymbol{x}_s - \boldsymbol{m}^f)^T (S^f)^{-1} (\boldsymbol{x}_s - \boldsymbol{m}^f)}. \tag{4.5}$$

The calculation of the adjusted color-related pixel features is conducted on the color-related image characteristics. They contain the pixel's color relative to the skin color distribution $(\boldsymbol{\mu}^{skin}, \Sigma^{skin})$, the brow color distribution $(\boldsymbol{\mu}^{brow}, \Sigma^{brow})$ and the lip color distribution $(\boldsymbol{\mu}^{lip}, \Sigma^{lip})$ of the current image. They are again represented by the Euclidean distance g^f and by the Mahalanobis distance h^f to the mean $\boldsymbol{\mu}^f$ of color distribution of the facial component f, see Equations 4.6 and 4.7 . Table 4.3 provides an overview of all extracted features. The calculation of the adjusted pixel features corresponds to Step 3 in Figure 4.2.

$$g^f = |\boldsymbol{x}_c - \boldsymbol{\mu}^f| \tag{4.6}$$

$$h^f = \sqrt{(\boldsymbol{x}_c - \boldsymbol{\mu}^f)^T (\Sigma^f)^{-1} (\boldsymbol{x}_c - \boldsymbol{\mu}^f)} \tag{4.7}$$

4.4.4 Regional Pixel Features

The idea of regional pixel features is to contribute information about the surrounding of a pixel. The core idea is that pixel in a certain area share common properties and that observing which properties are shared, provides information on the type of the pixel. For instance, the eyebrows are usually darker than the surrounding skin and therefore pixels located in a dark area in the upper face region are probably depicting eyebrows.

To integrate this information, we calculate an intermediate image that depicts the distance of every pixel's color value from the estimated skin color of the image characteristics. Intuitively, this provides an estimation of a skin color image and facial components, such as the mentioned eyebrow, will be determinable by their

a) Adjusted pixel features (color information)	
g^f	The Euclidean distance between the pixel's color and the mean color value $\boldsymbol{\mu}^f$ of the facial component f that has been determined for the entire image. It is computed as in Equation 4.6.
h^f	The Mahalanobis distance between the pixel's color and the mean color value $\boldsymbol{\mu}^f$ of the facial component f. It is computed as in Equation 4.7.

b) Adjusted pixel features (spatial information)	
$b^f_{\boldsymbol{x}}$	The probability of the pixel $\boldsymbol{x}$ to be part of the facial component f.
k^f	The Euclidean distance between the pixel's location and the predicted center $\boldsymbol{m}^f$ of the facial component f. It is computed as in Equation 4.4.
ℓ^f	The Mahalanobis distance between the pixel and the center of the facial component f. It is computed as in Equation 4.5.
i_l	The distance between the pixel location and a landmark position $\boldsymbol{x}_l$.

Table 4.3: Color and spatial information of the adjusted pixel features.

larger distance to the skin color estimation. For each pixel, we calculate a set of Haar-like features with the pixel in the center. Please note that the position of these Haar-like features depends on the pixel position within the image, which is part of the static image features. To link these feature to a specific position within the face region, the adjusted pixel features have to be considered. The feature size is linked to the ROI size and we extract features ranging for 0.03 to 0.3 ROI side size. Haar-like features consist of rectangular region, usually colored in white and black for visualization. They are specified by their position within the image, their scaling and their style, which refers to the arrangement of the regions. They are calculated by summing the pixel values in both regions and then subtracting one sum from the other. The feature styles used by our approach are depicted in Figure 4.6.

4.5 Classification

We train classifiers from extracted adjusted pixel features that determine the facial component depicted by single pixels. First, the face locater is applied to a set of training images in order to determine the ROIs. Then, in each image, the probability matrices are applied to calculate image characteristics. Finally, adjusted pixel features are sampled from all images. One classifier is trained per from the collected data per facial components to determine whether a single pixel depicts that specific facial component or not. We chose to train tree stumps boosted with AdaBoost, because this approach selects relevant features and reject less relevant features from the large number of provided features, which contributes to the execution performance of the algorithm. Training the classifiers corresponds to step 4 in Figure 4.2.

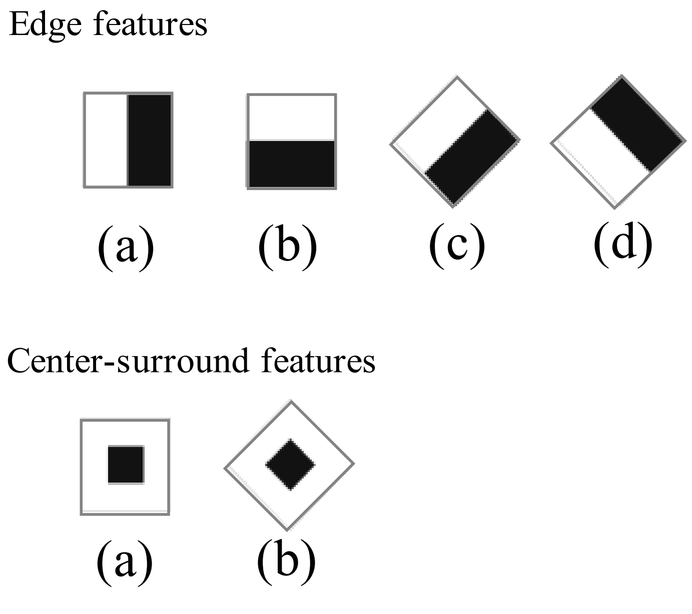


Figure 4.6: Haar-like features are extracted to describe the sorrounding of a pixel.

By applying a single classifier to all pixel of an image we obtain an estimation which pixels depict a specific facial component. We highlight these pixels in white to obtain a multi-band image $\boldsymbol{I}$. The underlying image-bands are denoted by $\mathcal{I}_f$. For instance, a multi-band image $\boldsymbol{I} = \{I_{skin}, I_{lip}, I_{brow}\}$ highlights face skin, lips and brows. Determining the multi-band image corresponds to step 5 in Figure 4.2.

4.6 Experimental Evaluation

Our evaluation is conducted on a data set of 376 face images that were collected from various Web pages. Each pixel is manually annotated with the facial component it depicts. The images have not been taken with a Computer Vision application in mind and therefore, the images include large variation with respect to pose, expression, age, gender, etc. We train our classifiers on 67% of the images and utilize the remaining 33% as test images to evaluate the accuracy of our approach. The training data is split, again, and half of it is utilized to generate the probability mask. The second half is utilized to train the facial component classifiers. The size of the probability masks is chosen to be $A^f \in \mathbb{R}^{100\times100}$. Our evaluation inspects the accuracy of each component separately: the probability masks, the image characteristics, and the classification of the facial components based on the adjusted pixel features.

4.6.1 Evaluation of the Probability Masks

The first evaluation inspects the reliability of the the probability masks entries. The evaluation is conducted on all images that were not utilized to generate the probability masks.

Applying a mask A^f to an image, each mask element $a^f_{i,j}$ covers a small number of image pixels, of which a certain fraction $\hat{a}^f_{i,j}$ depicts the facial component f. Our evaluation compares $\hat{a}^f_{i,j}$ to the mask element $a^f_{i,j}$. We compute the mean relative error $\hat{a}_{error}$ in this estimation according to Equation 4.8. Table 4.4 presents results for different facial components.

$$\hat{a}^f_{error} = \frac{1}{|ROI|} \sum_{ROI} \frac{|a^f_{i,j} - \hat{a}^f_{i,j}|}{a^f_{i,j}} \tag{4.8}$$

Since the size of the facial components differs, they cover a different area $|ROI^f|$ within the corresponding probability masks. Table 4.4 shows this area as well. Re-

facial component f	skin	lips	eyebrows	retina	sclera
mean error of all mask entries $\hat{a}^f_{error}$	3.86%	0.66%	0.64%	0.18%	0.18%
size of the mask [in pixels] $\lvert ROI^f \rvert$	5860	238	213	71	42
mean error of foreground mask entries $\bar{a}^f_{error}$	6.59%	27.73%	30.05%	25.35%	42.86%

Table 4.4: We evaluate the probability masks by inspecting the suggested probabilities on test images. The mean difference between the measured probabilities in the test images and suggested probabilities of the masks vary from 6.59% for skin and 42.86% for sclera.

lating the error of the entire mask to the area that a facial component covers yields the mean error of foreground pixels in this mask as given in Equation 4.9. According to Table 4.4, the skin color pixels are much more precisely estimated by A^{skin} than all the other facial component. The reason is that skin color covers a much more compact area within the face ROI and its shape is much less affected by facial expressions and head rotations than the area of the smaller facial components.

$$\bar{a}^f_{error} = \hat{a}^f_{error} \cdot \frac{|ROI|}{|ROI^f|} \tag{4.9}$$

4.6.2 Evaluation of the Image Characteristics

As mentioned in Section 4.4.2, the image characteristics model the spatial and color distribution of facial components. We evaluate the estimation of these distributions by calculating the relative error in the distribution parameters. Thereto, we determine distribution parameters from manual annotations in the test image and compare them with the distribution parameters gained by applying the probability masks.

To inspect the spatial image characteristics, we consider the facial component's estimated center location calculated from the manual labeling $\boldsymbol{m}^f_{manual}$ and the estimation $\boldsymbol{m}^f$ gained from the probability masks, normalized by the ROI side size. Equation 4.10 presents the computation of the error $\boldsymbol{m}^f_{error}$. Table 4.5 depicts this error for several facial components. For instance, the center location of the lips $\boldsymbol{m}^{lips}$ estimated by A^{lips} in test image is shifted by 0.06 of the side size of the ROI compared to its computed position from manual annotation.

facial component	lips	eyebrows	retina
$\boldsymbol{m}^f_{error}$ in ROI side size	0.066	0.059	0.062

Table 4.5: The estimated location of the center location of facial components is compared with manual annotations. Distances are given in ROI side size. Since this is roughly the face size, the distance estimation for all facial components shows an average error of below 8% of the face size.

$$\boldsymbol{m}^f_{error} = \frac{\boldsymbol{m}^f_{manual} - \boldsymbol{m}^f}{\boldsymbol{m}^f_{manual}} \tag{4.10}$$

The color-related image characteristics $(\boldsymbol{\mu}^f_{nRed}, \boldsymbol{\mu}^f_{nGreen}, \boldsymbol{\mu}^f_{nBlue})$ refer to the mean value for normalized red, green and blue color channel. Furthermore, the image characteristics consider their variances in $(\Sigma^f_{nRed}, \Sigma^f_{nGreen}, \Sigma^f_{nBlue})$. Again, we calculate their relative error and present it in Table 4.6. We observe that the estimation of $\boldsymbol{\mu}^f_{nRed}$, $\boldsymbol{\mu}^f_{nGreen}$ and $\boldsymbol{\mu}^f_{nBlue}$ is very accurate for skin, lip and brow in the test images, see Table 4.6. Therefore, these image characteristic form a robust basis for the computation of adjusted pixel features. However, the estimation of Σ^f_{nRed},Σ^f_{nGreen} and Σ^f_{nBlue}, which are important to calculate the Mahalanobis distance, are less robust. Please note that the Euclidean distance is not influenced by that. The estimation for the retina is the most inaccurate because these facial components occupy a small area of the face and are therefore more influenced by head movement.

color channel	$\boldsymbol{\mu}^f_{nRed}$	$\boldsymbol{\mu}^f_{nGreen}$	$\boldsymbol{\mu}^f_{nBlue}$	Σ^f_{nRed}	Σ^f_{nGreen}	Σ^f_{nBlue}
skin color	2.8%	3.7%	1.8%	41.4%	65.2%	74.8%
brow color	3.7%	5.2%	2.3%	44.5%	63.0%	70.7%
lip color	11.6%	7.0%	14.1%	48.9%	52.2%	59.2%
retina color	16.5%	20.4%	13.7%	208.3%	245.4%	437.6%

Table 4.6: Due to the small errors in the parameters of the skin, lip and eyebrow distribution estimation, they are utilized to compute the color-related adjusted pixel features.

4.6.3 Classifying Facial Components Using Different Features

This experiment inspects the impact of provided adjusted pixel features on classification accuracy. We train five example classifiers per facial component (C_{static}, $C_{color\ adj.}$, $C_{spatial\ adj.}$, $C_{all\ adj.}$, $C_{regional}$) and provide the classifiers with different sets of features. C_{static} is trained with the color information of the pixel only. It represents the traditional approach of considering the pixels' static color information only. $C_{color\ adj.}$ additionally takes the color information of adjusted pixel features into account. $C_{spatial\ adj.}$ uses the static color features and adjusted spatial features. $C_{all\ adj.}$ considers all features of the previously introduced classifiers. Finally, $C_{regional}$ additionally takes the regional image features into consideration. Please see Table 4.7 for a complete overview of the features provided.

Our evaluation applies the five classifiers to all pixels within the test images' ROIs. Table 4.8 and Table 4.9 illustrate the accuracy of each classifier. In Table 4.8 the number of correctly classified pixels (true positives and true negatives) is divided by the total number of pixels within the ROI. In Table 4.9 the number of pixels correctly classified as depicting the facial component is divided by the ground truth number of pixels depicting that facial component.

Please note that the number of pixels reflecting a facial component is only a small part of the complete region of interest and there are more negative examples than positive examples. Therefore, we down-sample the negative examples to have a comparable number of training and test examples for both classes.

As Table 4.8 indicates, $C_{regional}$ obtains the highest accuracy and clearly outperforms the other classifiers. It may be surprising that classifying skin shows the lowest accuracy in comparison to the other facial components. However, this is expected since skin covers a large area of the face and its appearance includes the

	static pixel features		adjusted pixel features		
	color	space	color	space	regional
C_{static}	provided	–	–	–	–
$C_{color\ adj.}$	provided	–	prodived	–	–
$C_{spatial\ adj.}$	provided	–	–	provided	–
$C_{all\ adj.}$	provided	–	provided	provided	–
$C_{regional}$	provided	–	provided	provided	provided

Table 4.7: The five classifiers of our evaluation consider a different set of features. None of them uses the static coordinate values, because they do not bear any useful information.

facial component	C_{static}	$C_{color\ adj.}$	$C_{spatial\ adj.}$	$C_{all\ adj.}$	$C_{regional}$
skin	79.6	86.1	84.9	89.6	90.3
brow	67.9	75.8	88.5	93.2	93.6
lip	84.8	92.0	95.4	96.5	97.0
retina	81.8	83.8	95.3	95.4	96.1

Table 4.8: The accuracy of classifying facial components significantly rises considering the adjusted pixel features.

facial component	C_{static}	$C_{color\ adj.}$	$C_{spatial\ adj.}$	$C_{all\ adj.}$	$C_{regional}$
skin	83.3	87.9	83.9	90.6	90.1
brow	52.8	67.4	81.6	92.5	91.7
lip	80.6	90.4	93.8	96.2	96.8
retina	88.5	86.5	95.9	96.7	95.4

Table 4.9: This Table inspects the fraction of pixels depicting a specific component that are recognized by the classifier.

larges variations due to shadows, one-sided lighting, specular points etc. Similar results are observable with classifying eye brows, but in this case varieties are induced by facial hair partly overlapping the forehead and the brows, rendering a clear separation between brows and environment difficult. These results illustrate that classifiers considering the adjusted features outperform static approaches. In general, providing the spatial-related features shows greater impact than providing the color-related features. Some example classification results are visualized in Figure 4.8 at the end of this chapter.

We utilize the chi-square test of significance to determine whether the influence of features is significant. Unfortunately, this is not the case for all classifiers with $C_{all\ adj.}$and $C_{regional}$. Furthermore, there was no significant difference between $C_{color\ adj.}$and $C_{spatial\ adj.}$for the skin classifier.

4.6.4 Evaluation of Face Locater Robustness Dependency

In this experiment, we inspect the impact of the face detection accuracy by the face locater on the feature extraction. We randomly move and scale the detected face region in the image, changing position and size at the same time. The magnitude of change is normalized by the size of the ROI. We move the located eye points with the ROI, but do not change their distance. The accuracy is calculated from the

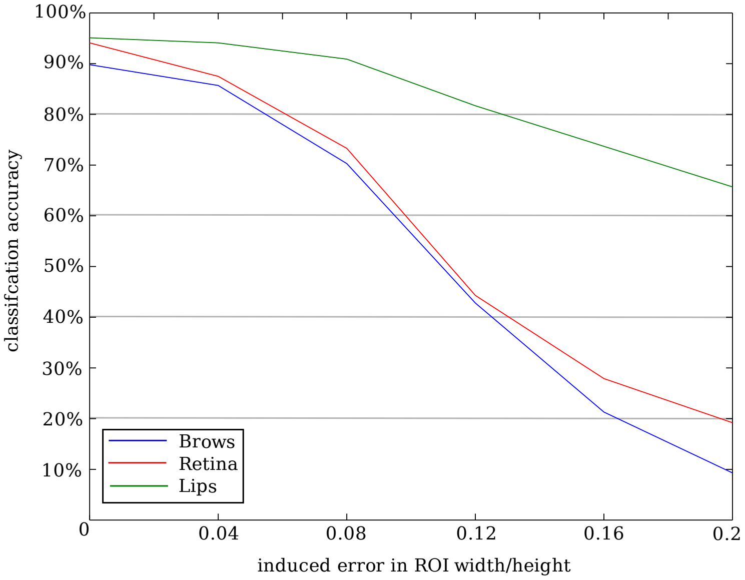


Figure 4.7: The degree to which the face locater accuracy influences the classification of the various facial components varies because the classifiers rely on different adjusted features.

fraction of correctly classified pixels, similar to the results presented in Table 4.8

As Figure 4.7 demonstrates, the degree to which the facial component classification is influenced varies greatly. Since the classifiers rely on the extracted features, the change in the classification accuracy reflects the feature robustness. The classification of the lips benefits from color-related features that are influenced less by the error induction, especially if depending on the skin color distribution. The classification of brows and retina, in contrast, relies on spatial-related features and therefore is heavily influenced by the induced error.

4.7 Discussion

In this section, we introduced pixel-based classifiers to segment the face from the background and facial components from the rest of the face. The advantage of our pixel-based classifiers is their balancing between speed and robustness. More complex approaches determine facial components region-based or fit a model to the face or face components. Although these approaches are usually more robust to image noise and context conditions like lighting or makeup, they are also computationally more expensive. Furthermore, we consider this approach as a preprocessing to subsequent model fitting rather than an alternative to model fitting. However, approaches that determine a pixel label only from color statistics, an approach typically applied to skin estimation, are usually faster due to our pixel feature adaption process. On the other hand, these approaches are not able to handle as large changes in lighting as our approach does.

Furthermore, although we applied our approach on face segmentation, it is not limited to this area. It is applicable to any task where the object of interest follows some characteristics in shape or texture that the classifier is able to exploit. Two requirement have to be met: It is required to provide annotated training images, which is usually a straightforward process. Furthermore, a detector for the object must be available, to restrict the search space and calculate the image characteristics. Other applications considered might involve the detection of license signs on cars, buildings on a landscape or segmentation of animal body parts.

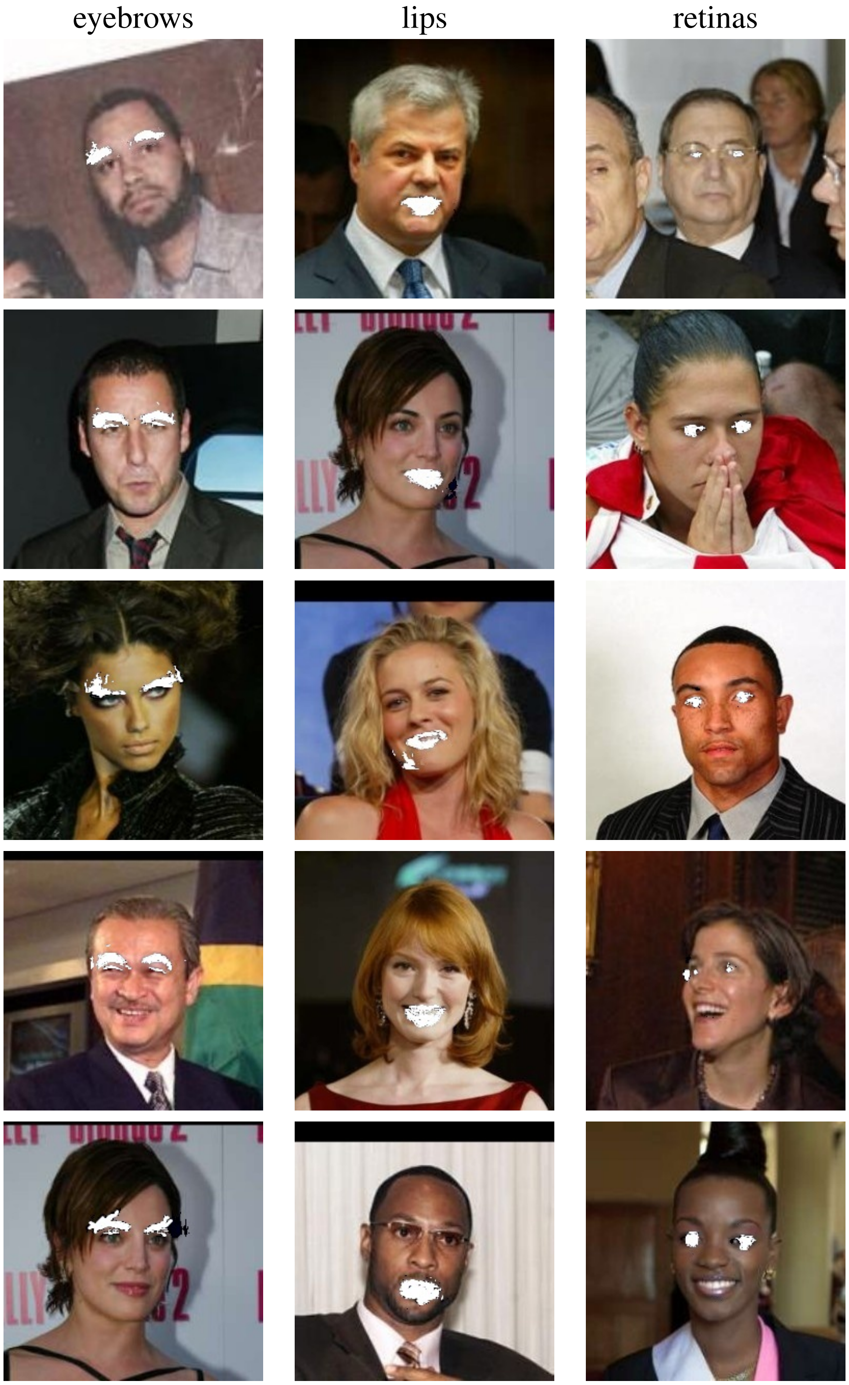


Figure 4.8: Image bands highlight the position of facial components in the image. Here, the original image data and different image bands are depicted in the same picture for demonstration.

Chapter 5

Model Fitting

The human face is a very important tool in everyday human communication. Humans obtain a lot of information from their fellow humans' faces, for example their identity, their emotional state via the facial expression or the focus of attention via the gaze direction. Therefore, the interpretation of human face images is a traditional topic also in computer vision research. The analysis of human faces provides information about person identity [137], facial expression [99] or head pose [90]. Face models, which represent the depicted face in a small number of descriptive model parameters that are collected in a parameter vector, are often applied for such tasks. However, an important step in doing so is to determine model parameters that match the image content without prior knowledge about the face visible in the image. This process is referred to as ”face model fitting” or ”face alignment”. In this chapter, we present our approach to take this step, see Figure 5.1 for some example images, on which a face model has been fit automatically.

The first challenge to be solved is the detection of faces within the image, a task also sometimes referred to as *face location* or *face localization*. Human capability in this regard are very high, as they are able to compensate for the effects of partial occlusion by perceiving the face as a whole, with visual imagination covering for the lack of visual perceivability. With regard to images of actual faces, it is a common assumption that the threshold for the detection of faces within images by human observers is at 100 to 200 pixels consisting of two gray-levels [109, 14]. The human brain is trained to actually recognize patterns of faces where in reality none exist, such as within clouds, a phenomenon know as pareidolia [7]. To perform this step automatically, different approaches have been suggested, ranging from IR-illumination of the eyes to sliding window-based approaches or color statistics [53, 60, 136, 50]. The approach of Viola et al. is often relied on, since it is known for its robustness and speed [136]. Face detection constitutes a challenge that becomes especially tough whenever vital parts of the face are occluded. Both, head pose and scenery, might lead to occlusion of facial features, up to the point where recognition of the face becomes impossible. Automatic approaches to face detection are usually based on the detection of a certain fixed pattern in the image, e.g. finding salient facial features like eyes, nose and mouth whose geometric relation towards each other is known, and are therefore often much less robust as far as partial occlusion of the face is concerned.

The second challenge is obtaining a numerical representation of the face structure, which is the main focus of this chapter. Although the general structure of human faces is independent of gender, age or ethnic background, human faces vary from one person to another with respect to (face) size, facial hair, hair color,

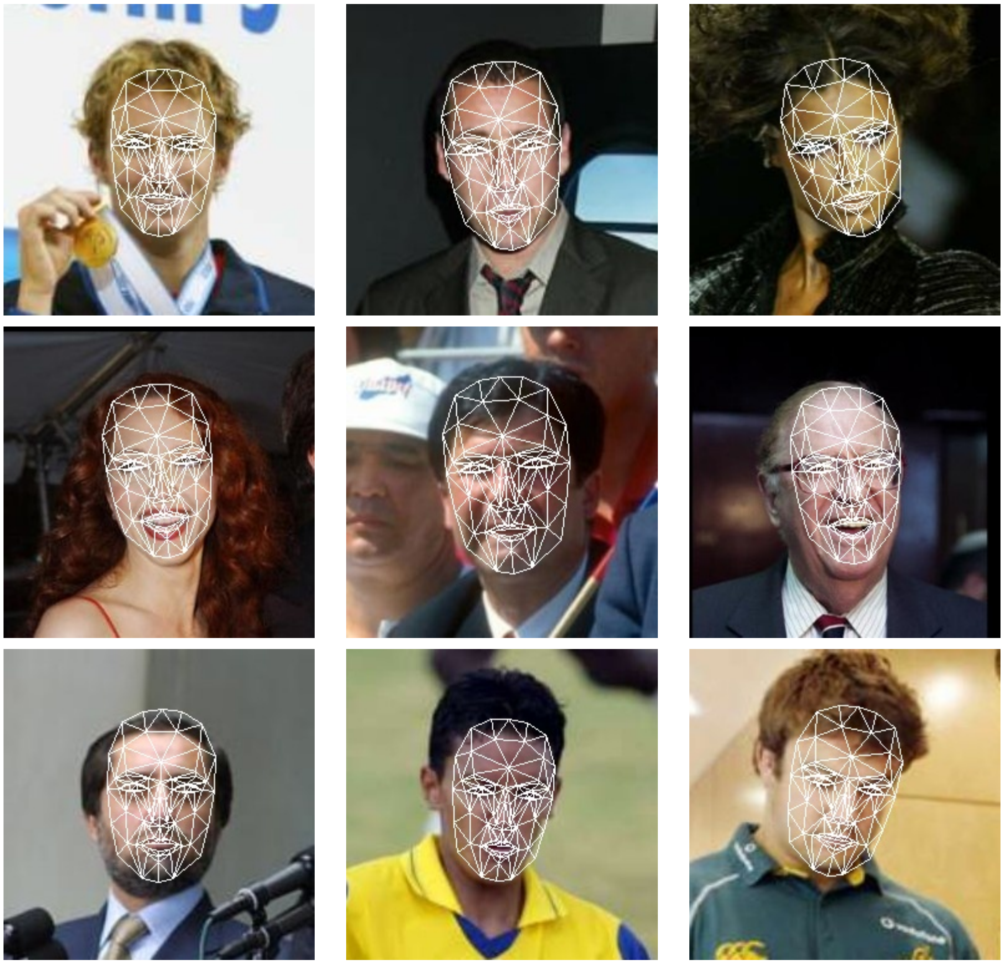


Figure 5.1: We fit a face model on image of the ”Labeled Faces in the Wild” database [54].

wrinkles or complexion. Furthermore, the specific appearance of a face is, as in any vision application, influenced by context conditions like lighting or head pose. Therefore the raw image data offers large varieties, even if only the face region is inspected. To handle these varieties, face models provide an abstract representation of the image content. Depending on the application in mind, they consider the position of specific landmarks in the image [22], the face texture [9, 18], the face pose in 3D space [9, 2] or the 3D face structure [18] in their parameters. Often, face models are created by calculating statistics on a large number of training images [9, 18, 20, 22].

Since our application in mind is facial expression recognition, we apply a three-dimensional face model that considers face pose and face shape in 3D space. Although approaches that additionally take the face texture into consideration are usually more robust, they also require a magnitude of additional calculation time, which currently prevents them from reaching real-time capability in 3D space. Therefore, instead of integrating texture information in the face model, we provide the fitting algorithm not only with the raw image data, but with multi-band images that have been obtained as described in Chapter 4.

5.1 Problem Statement

Face models and related model fitting strategies determine landmarks with a specific semantic interpretation, such as the corners of the eyebrows, eyes, lips, or the face contour. However, most face models obtain the dependencies between these points from statistics in a set of training images, and thus model parameters do not refer to any semantic interpretation. One example for extracting these dependencies is learning them via Principal Component Analysis (PCA). Therefore, for instance to infer whether a person's mouth is opened, statistical or rule-based methods have to be applied, since this information is not directly available in a single model parameter. The reason is that model parameters refer to face appearance in the image rather than the real-world face state. Unfortunately, such information is very important for our target application. Furthermore, most models are 2D models and do not represent the face shape in 3D space but in image coordinates.

5.2 Solution Idea

Therefore, we chose to integrate the Candide-III face model, which is publicly available, offers parameters with semantic interpretations and represents the face structure in 3D [2]. Its model parameters refer to standard face descriptors, like the FACS [34] or the mpeg4 standard. Since face model fitting algorithms are often tuned towards a specific face model, this also calls for a novel face model fitting approach, which is specifically tuned to this model. We will provide such an approach with the integration of multi-band images. Since recent research has demonstrated the superiority of displacement expert based over objective function based approaches, we will construct displacement experts for fitting.

5.3 Related Work

Face models have proven to be a powerful tool in computer vision for a number of face image analysis applications. In this section, we will present the most important face models available to the community. Usually, model fitting algorithms are tied to a single model. Therefore, we will also introduce the most important model fitting strategies.

5.3.1 Face Models

Face models are discriminable into modeling the face shape only or taking the face texture into consideration, as well. The advantage of models in the second category is that they can be rendered to reproduce the face appearance in the image, which provides application opportunities, like creating realistic faces in movies [10].

An example for the first category are Active Shape Models (ASMs), which have been proposed already in the early 90s by Cootes et al. [19]. They are generated by extracting statistics on manually specified landmarks in a set of training images, usually learned via PCA. Therefore, the model parameters are projected to a set of model points in the image. Several years later, the research group introduced the most famous representative of the second category, the Active Appearance Model (AAM) [18]. In addition to the face shape, the face texture is also integrated in the PCA.

Two further ideas emerged from AAMs: Firstly, Blanz et al. generate a 3D model of human faces from laser scans and colorize them with parameterizable texture [9]. This model is called ”3D Morphable Model”(3DMM) and is rendered with a lighting model taken from computer graphics to obtain highly realistic face images. An advantage of this model is that real-world information like the head pose or gaze direction is directly available without any further interpretation. Secondly, Cristinacce et al. propose representing the appearance of single templates around model points rather than the holistic face appearance and call this a ”Constrained Local Model” (CLM) [22]. This models does not inherently constrain relative model point locations, but consider them independent in their very basic formulation.

The Candide-III face model has been proposed by Ahlberg for image and video coding [2]. It models the face structure with 113 manually specified 3D vertices. Its parameter vector refers to established description methodologies of the human face, like the FACS or the mpeg-4 video codec standard.

Very recently, Wu et al. introduced the Boosted Ranking Model [147]. It represents face texture in Haar-like feature values instead of pixel color values and is fitted with boosted ranking. Therefore, although it takes the face texture into consideration, this model can not be rendered to the image. Please note that except for the 3DMM and the Candide-III face model, all models presented are 2D models and do not consider the real face structure or face pose but only the face shape or texture in the image.

5.3.2 Model Fitting

Most fitting strategies fall in one of two categories, utilizing either *objective functions* or *displacement experts*. These fitting strategies are sometimes also termed ”discriminative fitting” (displacement experts) and ”generative fitting” (objective functions) [116, 144]. Objective functions $\boldsymbol{f}(\mathcal{I}, \boldsymbol{p})$ yield a comparable value that determines how accurately a parameterized model $\boldsymbol{p}$ fits to an image $\mathcal{I}$, and are optimized to determine the optimal parameterization $\boldsymbol{p}_{\mathcal{I}}$. In contrast, displacement experts $\boldsymbol{g}(\mathcal{I}, \boldsymbol{p})$ propose a parameter update directly to calculate the optimal parameterization $\boldsymbol{p}_{\mathcal{I}} = \boldsymbol{p} + \boldsymbol{g}(\mathcal{I}, \boldsymbol{p})$. Since objective functions have been proposed earlier, and newly proposed face models are usually considered with objective function-based fitting first, there is more work available on objective functions than on displacement experts.

Objective Functions

Typical examples of objective function based fitting are presented with application to ASMs [19, 145] and to AAMs [18, 81, 93]. Objective functions for ASMs are computed from extracted image features around the model points. Early proposed fitting strategies used simple features and manually designed rules. For instance, the first proposed fitting simply considered the distance to the nearest image edge [19]. More complex features have been proposed by Romdhani et al. [108]. Since this has been proven to be not very robust, especially in the context of face models, Wimmer et al. propose to learn the objective function from annotated images instead of manually designing its calculation rules [145]. The idea is that annotating images with examples of well-fit face models is much more intuitive than designing the rules to obtain such a parameterization, and is therefore less error prone. Their approach is evaluated on the BioID database. Recently, Ding et al. proposed an approach that detects the face and several facial features in the image with a combination of machine-learned and manually designed calculation rules [28]. Although they consider this a detection approach, its idea is more related to what is considered model fitting in this thesis due to the level of detail and since the detection of facial components is restricted by previous face detection. The novelty of their approach is that they propose to include misaligned detection examples in the training data. Evaluation is conducted on the XM2VTS database.

Determining the fitness of AAMs is straightforward, simply by comparing the rendered face model with the original image data, which allows for a high accuracy given a good initialization. Since the formulation of the objective function is simple with AAMs, the question arises how this function is optimized in a fast and accurate way. Matthews et al. present a comprehensive survey on this topic and introduce a novel idea that has found much interest, since it provides a fast optimization [81]. Unfortunately, their evaluation is not conducted on a standard database, but on data that is made public on their homepage. However, a remaining major drawback of AAMs is that they tend to get stuck in local optima during fitting. Therefore, Nguyen et al. recognize the objective function rather than the optimization strategy to be the weakness of this strategy [93]. They propose to learn the objective function in order to fulfill properties desirable for optimization. They aim at increasing the convergence radius and accuracy, similar to the approach that has been proposed by Wimmer et al. for ASMs. Evaluation is presented on the CMU Multi-PIE database. The typical fitting of 3DMMs is also based on rendering the model to the image and computing an error image [9, 8]. The disadvantage with these models is that the fitting is computationally very expensive and a good initial estimation of the model parameters is required. To obtain a larger region of convergence, Hamsici et al. extend AAMs and 3D Morphable Models with rotation invariant kernels that are fit applying the kernel-trick [47]. A similar approach is presented by Kemelmacher et al. [64].

Displacement Experts

Displacement expert-based fitting for AAMs has been proposed by Saragih et al. [116]. They train boosted regressors on Haar-like features from the error image to determine parameter updates. Comparing it to the objective function-based approach of Matthews et al. [81], they recognize their approach comparable in terms of speed and significantly superior in terms of accuracy when tested on the XM2VTS database. In their subsequent work, they present a sophisticated displacement expert training strategy that simulates the model fitting on the training data. This allows for adapting the update function to the fitting iteration and circumvents the requirement to compute the complete error image [115]. In addition to the XM2VTS database, they also include the CMU Multi-PIE database for evaluation. Their former strategy is outspeeded by this novel idea. Similar conclusions are drawn by Cristinacce et al., who compare an objective function-based fitting with a displacement expert-based fitting for ASMs on the BioID database [23]. They train local detectors for each model point with slightly different behavior: Their first approach is to train classifiers, which decide whether a single model point is well-fit at a specific location and determine the fitting via exhaustive search. Their second approach is to train local regressors, which provide a position update directly. This already very much reminds of CLMs and therefore it might have inspired the work of Wang et al., who derive CLM fitting based on displacement experts. They compare them to exhaustive search and experience a significant gain in speed and accuracy, tested on the Multi-PIE database [151]. We refer to our earlier work for a preliminary version of our current work [82].

Recent Work

CLMs are fit by evaluating response maps for every model point, which state the probability that this model point is well-aligned at a specific image location. Fitting these models is mainly concerned with determining, which position results in the highest probability. Often, parametric approximations of these response maps are calculated for this task. A recent work by Saragih et al. provides a short survey and proposes a different approach that constructs a displacement expert from a nonparametric representation of response maps [118]. They also consider the idea to utilize multiple local experts for fitting, an idea that is applicable to other fitting strategies as well [117]. Evaluations are conducted on the CMU Multi-PIE and the XM2VTS database.

An image based method that is neither objective function-based nor displacement expert-based, since it does not rely on a classical face model, is presented by Zhu et al. [156]. Instead of fitting a face model to the face, they compute a warping of the image face region to a template face. The downside of this approach is that no specific information about face structure, like mouth opening, is available. On the other hand their algorithm shows great accuracy with respect to the image registration. They compare their method with the method presented by Gu et al. on example images taken from the ”Labeled Faces in the Wild” database [152]. They show that Gu’s method has severe inaccuracies when confronted with nonfrontal images, which is mainly to the fact that the method has been trained and evaluated on frontal faces only. Gu et al. propose a fitting strategy for a shape model that relies on the expectation-maximization algorithm [152]. They assign weights to single model points that reflect the estimation robustness, which allows for handling even large occlusions. They evaluate their approach on frontal-view images specifically considering large occlusions and strong facial expressions.

Candide-III Model

Since we explicitly evaluate our approach on the Candide-III model, we will review some of the published work considering this face model separately. Dornaika et al. utilize it for face tracking and simultaneous facial expression recognition [26]. However, the model fitting is conducted manually. They add this missing component in their subsequent work and provide an automatic texture-based fitting [29]. However, their texture model is person-specific and requires person-specific and manually annotated training data. Unfortunately, no evaluation on standard database data is conducted. Another automatic fitting approach is proposed by Sheng et al. [123]. They extract several facial landmarks in the image, like the eyebrow corners, and approximate model parameters so that model points match the landmark positions. However, comparable evaluations are not provided. Kotsia et al. utilize the Candide-III face model for facial expression recognition on a standard database but fit the model manually, too [67]. Chen et al. present a model tracking approach for the Candide-III face model, which requires only a single annotated frontal image of the person for training [15]. However, they also obtain this annotation manually. All mentioned approaches rely on a subset of the full parameter set only, which are selected to suit the specific application.

Multi-band Image Based Fitting

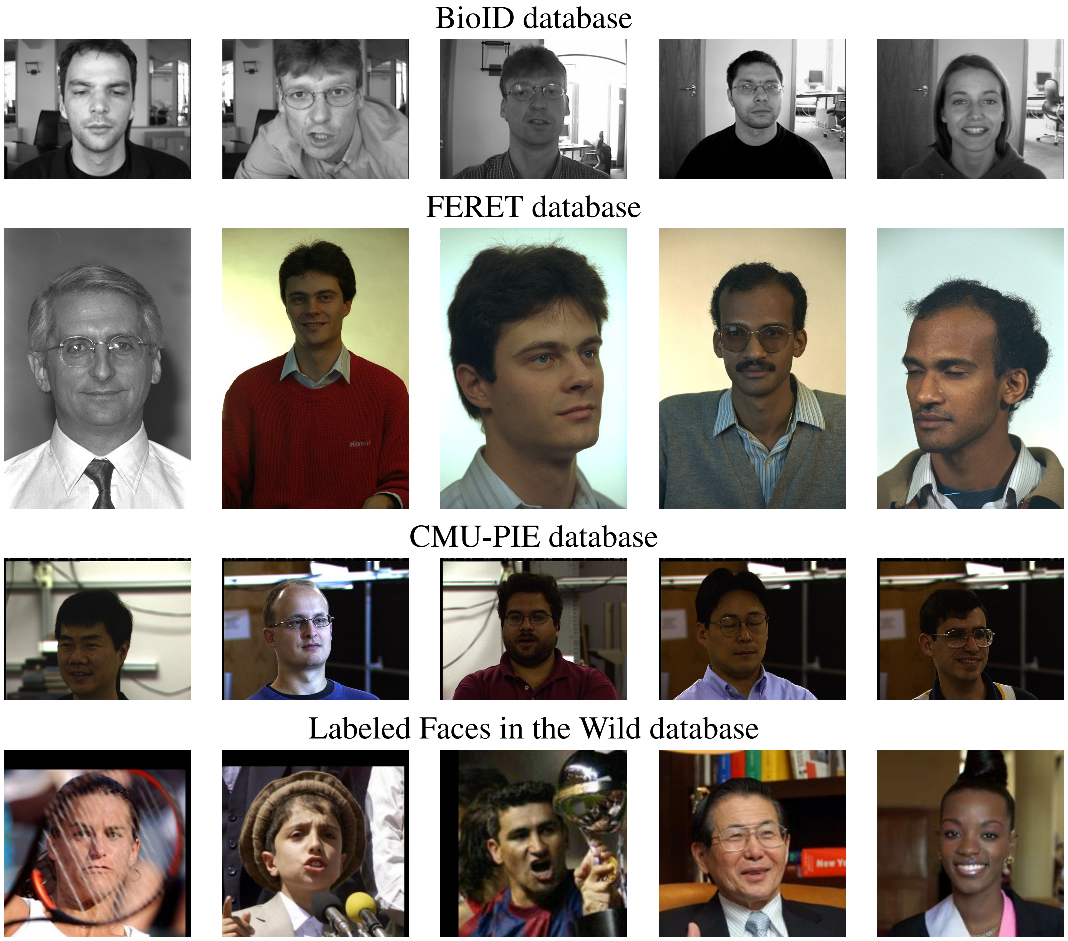
Cootes et al. [20] propose to utilize images with two image bands for creating and fitting face appearance models. These image bands reflect edge directions in two dimensions, where the magnitude indicates the degree of reliance in the orientation estimation. Therefore, the appearance model is not rendered as intensity values but as edge directions. This approach is similar to ours because not only the raw image data but an image representation with various additional image bands is considered. Similarly, Stegmann et al. [128] propose to utilize a multi-band image representation instead of raw intensity values for fitting active appearance models. Their so-called VHE image considers conversions of the original image data in three different color spaces. They experience a significant gain in accuracy. Kahmaran et al. [62] also follow this idea but rely on a different color conversions. In contrast to these approaches, our representation adapts to image conditions and the characteristics of the visible person. Similar approaches are described in [23, 144, 145].

5.3.3 Conclusions

Usually, approaches are evaluated on standard image databases that restrict the image content with respect to background, lighting, head pose or facial expression. Most image data is captured in controlled lab or office environments. The BioID database offers 1521 gray-scale images from 23 subjects that are taken inside an office environment [57]. The XM2VTS database depicts 295 people in 1180 images, therefore the person variability is much higher [86]. However, the background in most images is very neutral and the images are also taken in controlled lighting. The FERET database includes images with controlled head rotation and facial expressions to increase the data variability [101]. A similar approach is taken in the assembly of the CMU Multi-PIE database [45], the successor of the CMU PIE database [130], but lighting changes are induced, additionally. However, all these database consist of images that have been captured with a computer vision application in mind in a structured environment. Although, artificial induced illumination and head pose changes create challenging data, they do not necessary reflect the variability of real-world scenarios. For instance, most images are taken with a plane background. Artificially induced variations again follow a pattern based on author assumptions and expectations. Figure 5.2 depicts some example images from several databases.

In contrast, we propose to evaluate face model fitting algorithms in an unconstrained environment. We use the ”Labeled Faces In The Wild” database, which contains images that have been taken in real-world conditions [54]. These images depict persons of public life and have been collected from the media, spanning a large variety of ethnic backgrounds, age, facial expressions, lighting and image backgrounds.

Comparison of displacement expert based fitting and objective function based fitting leads to the conclusion that displacement experts are faster and more robust [116, 23, 115]. The reason for their superiority in speed is that they have to be evaluated less often due to their basic formulation. The reason for their superiority in accuracy is their robustness against local optima. The disadvantage is, however, that formulating the objective function is often easier. Therefore, we rely on displacement experts for fitting, but provide a short comparison to objective functions based fitting, as well.



BioID database

FERET database

CMU-PIE database

Labeled Faces in the Wild database

Figure 5.2: Some example images taken from four databases. The images are chosen to reflect the variety in the databases's image content. Unfortunately, no examples of the XM2VTS database may be depicted here, since it is not free of charge.

5.4 System Overview

The goal of face model fitting is to calculate model parameters $\boldsymbol{p}_{\mathcal{I}}$ that approximate the *preferred* model parameters $\boldsymbol{p}^*_{\mathcal{I}}$ for a given image $\mathcal{I}$. These preferred model parameters are usually specified manually, assuming that the human annotator provides ideal model parameters. Therefore, a perfect model fitting algorithm would result in $\boldsymbol{p}_{\mathcal{I}} = \boldsymbol{p}^*_{\mathcal{I}}$. Fitting starts with an initial parameter estimate

$\boldsymbol{p}$. Although $\boldsymbol{p}$ can be chosen arbitrarily, we propose to apply a face detector to calculate an initial estimation of the model pose. As has been mentioned already, there are two major fitting strategies, utilizing either objective functions or displacement experts. Both approaches are formalized in Equations 5.1 and 5.2, respectively. We denote displacement experts as $\boldsymbol{g}$ and objective functions as $\boldsymbol{f}$. An important difference between these approaches is that the displacement expert is evaluated only once, whereas the objective function is usually evaluated more often, depending on the implementation of the $argmin$ operator. Therefore, displacement experts consider only a single model parameterization whereas objective functions inspect a large number of model parameterizations. Figure 5.3 visualizes the application of a displacement expert and an objective function for fitting.

$$\boldsymbol{p}_{\mathcal{I}} = \boldsymbol{p} + \boldsymbol{g}(\mathcal{I}, \boldsymbol{p}) \qquad \text{displacement expert} \qquad (5.1)$$
$$\boldsymbol{p}_{\mathcal{I}} = \underset{\boldsymbol{p}}{argmin}(\boldsymbol{f}(\mathcal{I}, \boldsymbol{p})) \qquad \text{objective function} \qquad (5.2)$$

Please note that the output of $\boldsymbol{f}(\mathcal{I}, \boldsymbol{p})$ is always a scalar value whereas the output of $\boldsymbol{g}(\mathcal{I}, \boldsymbol{p})$ is in the domain of model parameters. The accuracy of objective function-based fitting approaches depends on the accuracy of $\boldsymbol{f}$ and on the accuracy of the optimization algorithm. In contrast, with displacement experts the accuracy of the fitting solemnly depends on the accuracy of $\boldsymbol{g}(\mathcal{I}, \boldsymbol{p})$. To increase the accuracy of the underlying functions $\boldsymbol{f}$ and $\boldsymbol{g}$, we propose to provide them not only with the original image $\mathcal{I}$ but with a multi-band image representation $\boldsymbol{I}$ as shown in Equations 5.3 and 5.4, consisting of the image-bands $\boldsymbol{I} = \{\mathcal{I}, I_{skin}, I_{lip}, I_{brow}, I_{retina}\}$.

$$\boldsymbol{p}_{\mathcal{I}} = \boldsymbol{p} + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p}) \qquad \text{displacement expert} \qquad (5.3)$$
$$\boldsymbol{p}_{\mathcal{I}} = \underset{\boldsymbol{p}}{argmin}(\boldsymbol{f}(\boldsymbol{I}, \boldsymbol{p})) \qquad \text{objective function} \qquad (5.4)$$

Basically, a displacement expert can always be converted to an objective function, because the parameter update is essentially a parameter error and the displacement expert an error function. Since most optimization strategies require a real-valued function, the displacement expert output has to be converted to real values, for instance by computing $|\Delta\boldsymbol{p}|$. The definition of an objective function is straightforward: $\boldsymbol{f}(\mathcal{I}, \boldsymbol{p}) = |\boldsymbol{g}(\mathcal{I}, \boldsymbol{p})|$. This approach has been taken in Figure 5.3,

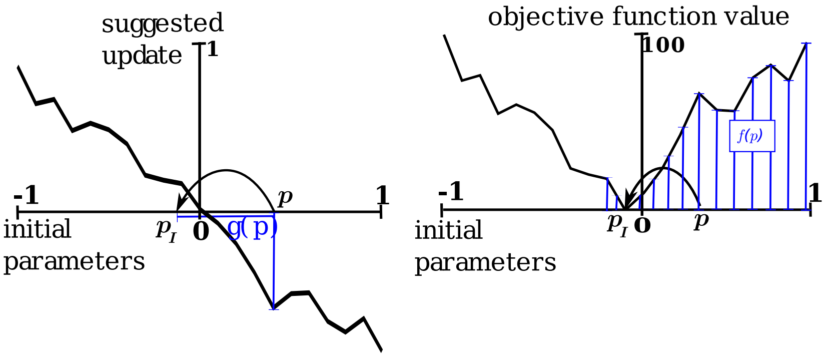


Figure 5.3: Displacement experts are evaluated only once, since they directly propose a parameter update. Objective functions are evaluated more often, here, visualized with search-based optimization. Blue lines indicate function evaluations.

even if in this example the displacement experts also resulted only in a single parameter value. Please note that a displacement expert can not be obtained from an objective function. For this reason, and because displacement experts are much faster in their execution, we will mainly focus on the generation of displacement experts in the remainder of this chapter and derive objective functions from them, if required.

We apply our fitting strategy on the Candide-III face model. This manually designed 3D face model consists of $K = 116$ model points that are arranged in 184 triangles [2]. Please see Figure 5.4 for some visualized example parameterizations. It does not inherently include face texture, which allows for a faster fitting. Model parameters specify relative model point positions to influence the model shape. Vertex coordinates are calculated by applying the shape deformation $\boldsymbol{v}_{shape} = \boldsymbol{S}\boldsymbol{s} + \boldsymbol{A}\boldsymbol{a}$ and a scaling factor $\boldsymbol{c}$ to the basic model structure $\boldsymbol{v}_{basic}$. The difference between the parameters in $\boldsymbol{s}$ and $\boldsymbol{a}$ is that $\boldsymbol{A}$ contains motion that may appear due to facial expressions whereas $\boldsymbol{S}$ contains vertex motions to adapt the general face structure to the face structure of a specific person. The advantage of this model is that its parameters model face shape changes with semantic meaning, which renders it highly suitable for facial expression recognition. However, its drawback is that some parameters describe very similar face deformations with only little semantic difference, like the $Upper\ lid\ raiser$ and the $Eye\ close$ parameter. Furthermore, some parameters model face deformations that are very difficult to detect, even for humans, like the $Cheek\ Z - extension$ parameter. Finally, a rotation matrix $\boldsymbol{R}$ and a translation $\boldsymbol{t}$ specify the model pose. The $3K$

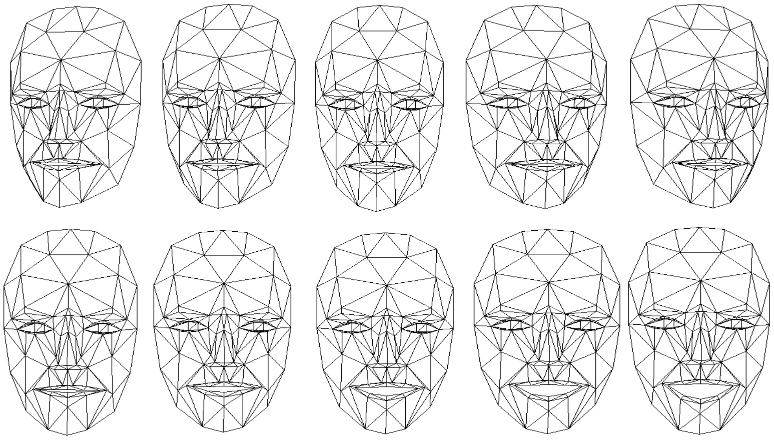


Figure 5.4: Model parameters change point positions to reflect face pose or shape change

dimensional vector $\boldsymbol{v}$ contains the vertex x-, y- and z-coordinates. In total, the model vertex coordinates are computed according to Equation 5.5. Table 9.2 in the Appendix summarizes the model parameters considering facial expressions and the face shape.

$$\begin{aligned}\boldsymbol{v} &= \boldsymbol{cR}(\boldsymbol{v}_{basic} + \boldsymbol{v}_{shape}) + \boldsymbol{t} \\ &= \boldsymbol{cR}(\boldsymbol{v}_{basic} + \boldsymbol{Ss} + \boldsymbol{Aa}) + \boldsymbol{t} \end{aligned} \tag{5.5}$$

We denote the single vector elements, i.e. single parameters by p_i. The parameter vector is assembled according to Equation 5.6.

$$\boldsymbol{p} = [\boldsymbol{c}, \boldsymbol{t}^T, r_x, r_y, r_z, \boldsymbol{a}^T, \boldsymbol{s}^T] = [p_1, .., p_n, .., p_N] \tag{5.6}$$

5.5 Training Displacement Experts

In this section, we describe our proposed approach to train displacement experts from manually specified annotations. Our approach is based on the idea of a perfect displacement expert, which always determines the correct parameter update. Therefore, applying it always results in a perfect model fit $\boldsymbol{p}^*_{\mathcal{I}}$. Obviously, if the correct model parameterization $\boldsymbol{p}^*_{\mathcal{I}}$ is known, such a displacement expert is easy to construct, as Equation 5.7 depicts. Unfortunately, the ideal model parameterization is usually not known, unless it is specified manually, which prevents this approach from being practically applied. However, it will be applied to generate training data to train a further displacement expert $\boldsymbol{g}^{\ell}(\boldsymbol{I},\boldsymbol{p})$ that is independent of $\boldsymbol{p}^*_{\mathcal{I}}$, as presented in Equation 5.8.

$$\boldsymbol{g}^*(\boldsymbol{p}^*_{\mathcal{I}},\boldsymbol{p}) = \boldsymbol{p} - \boldsymbol{p}^*_{\mathcal{I}} \tag{5.7}$$

$$\boldsymbol{g}^{\ell}(\boldsymbol{I},\boldsymbol{p}^* + \Delta\boldsymbol{p}) = \Delta\boldsymbol{p} = \boldsymbol{g}^{\ell}(\boldsymbol{I},\boldsymbol{p}) \tag{5.8}$$

5.5.1 Image Annotation

Since the perfect displacement expert relies on manually annotated model parameters, the first step is to provide $\boldsymbol{p}^*_{\mathcal{I}}$ in a set of training images. Unfortunately, this is a laborious step. However, it is the only step involving manual work, and several databases provide annotation with the image data like the BioID database [57]. Annotating a single image takes an experienced person $2-4$ minutes. Nevertheless, we capture a set of images ourself with changing facial expression, head pose, person identity and lighting. The reason is that annotating the images is easier when multiple views of the scene are visible. Therefore, we capture training data with a calibrated system of three cameras to obtain three different views of a single scene as demonstrated in Figure 5.5. Please note that these images are used only for training, not for evaluation. We integrate publicly available data for evaluation. In total, the dataset consisted of 87 scenes, which sums up to $3 \cdot 87 = 261$ images. Obviously, for all these images it is desirable to have $\boldsymbol{g}^{l}(\boldsymbol{I},\boldsymbol{p}^*_{\mathcal{I}}) = 0$, since no parameter update is required in this case. However, we also need training examples that reflect cases, when a parameter update is required.

5.5.2 Training Data Generation

This training data is acquired automatically, by inducing random model parameter variations $\Delta\boldsymbol{p}$ to obtain new model parameterization $\boldsymbol{p} = \boldsymbol{p}^*_{\mathcal{I}} + \Delta\boldsymbol{p}$. In later execution, the displacement expert will determine these induced parameter variations according to Equation 5.8.

Therefore, training data consists of pairs of images and example parameterizations $< \boldsymbol{I},\boldsymbol{p} >$ that are labeled with the induced parameter error $\Delta\boldsymbol{p}$. The displacement expert will be trained to perform a mapping $< \boldsymbol{I},\boldsymbol{p} > \rightarrow \Delta\boldsymbol{p}$, which

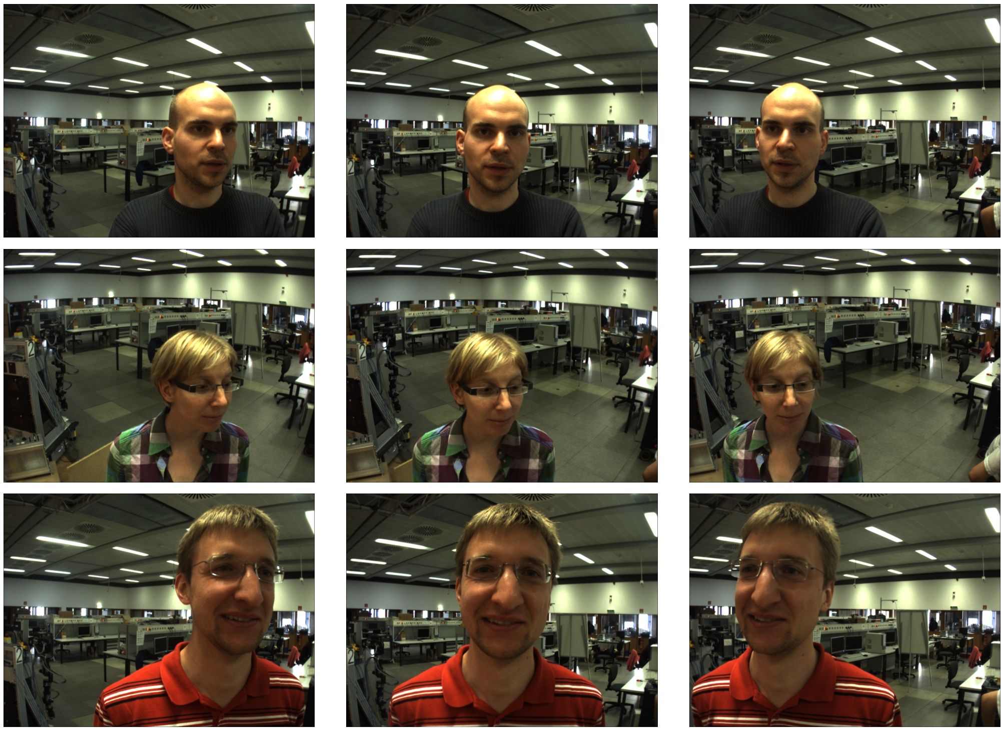


Figure 5.5: The face is visible in multiple camera images that are captured simultaneously. To the annotating person, the face position in 3D space is easier to estimate than in image data taken by a monocular camera.

allows to fit the model according to Equation 5.3. For training purpose, this collected training data can be thought of as being sampled from a perfect displacement expert. In order to simplify this learning problem, we train single displacement experts for each single parameter separately. Therefore, instead of training one displacement expert that determines an update for the complete parameter vector, we train a set of displacement experts, each proposing a parameter update for a single element of the parameter vector only and we refer to them as ”local displacement experts”. Equation 5.9 formalizes this step.

$$\Delta p_i = g_i^{\ell}(\boldsymbol{I}, p_i^* + \Delta p_i) \tag{5.9}$$

5.5.3 Feature Extraction

To train the displacement expert, it needs to be provide with a set of features that link the image annotation and the image content. Theoretically, a set of single pixel values or even the raw image data would provide this information. However, the learning algorithm will have a hard time generalizing well from such noisy data. Therefore, we extract descriptive low-level image features. In this thesis, we gather Haar-like features in different styles and sizes [72]. Please see Figure 5.6 for a the Haar-like features utilized in our approach.

As mentioned, model parameters change the relative positions of model points. However, most model parameters influence only a small subset of model points. For instance, the parameter that represents rising eye brows has no influence on model points at the chin. Therefore, we extract image features only in the neighborhood of model points influenced by a single model parameter p_i. Feature values are calculated by $h_i^k(\boldsymbol{I}, \boldsymbol{p})$ with k denoting the feature index. The feature index specifies, in which image band and style, and at which position and scaling a feature is extracted. Features are extracted at model point positions and at po-

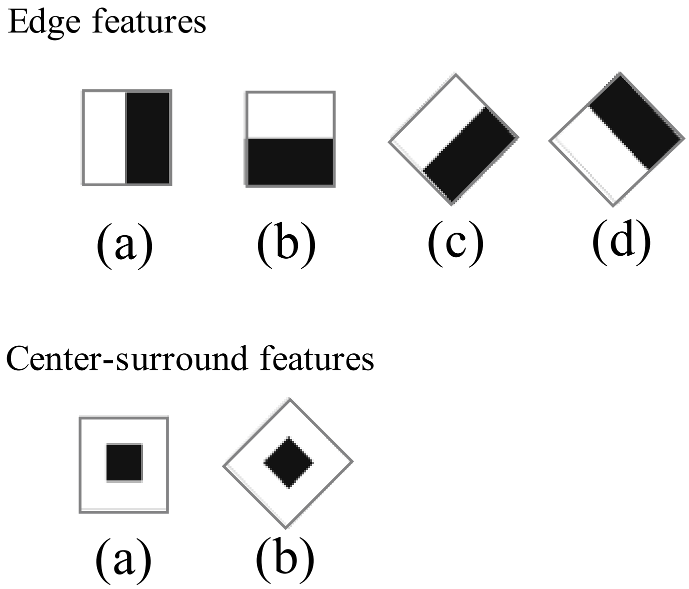


Figure 5.6: Haar-like features are extracted from the image as specified by the model parameters to link the image annotations to the image content.

sitions along the model point motion defined by the model parameter. We denote the number of features that are extracted for a single model parameter i by H_i. However, in the remainder of this section we will refer to H_i simply by H for improved readability of equations.

Usually, Haar-like features are calculated from single channel images to obtain numerical values. Therefore, we split our multi-band image to its underlying image bands and calculate feature values from every image band separately. The same set of positions, styles and scalings is used for the feature extraction in every image band. Please note that the amount of features varies with the number of model points influenced by the model parameter. We choose to extract Haar-like features at 3 different positions in 2 different sizes, in 6 different styles and from 5 different image bands, summing up to 180 features per model point. For instance, parameter p_7, the "Jaw drop" parameter, influences the spatial location of 10 model points, and we extract $H = 1800$ features $(h_7^1(\boldsymbol{I}, \boldsymbol{p}), h_7^2(\boldsymbol{I}, \boldsymbol{p}), ..., h_7^{1800}(\boldsymbol{I}, \boldsymbol{p}))$. All extracted features are assembled in an image feature vector $\boldsymbol{h}_i(\boldsymbol{I}, \boldsymbol{p})$ as shown in Equation 5.10.

$$\boldsymbol{h}_i(\boldsymbol{I}, \boldsymbol{p}) = (h_i^1(\boldsymbol{I}, \boldsymbol{p}), h_i^2(\boldsymbol{I}, \boldsymbol{p}), ..., h_i^H(\boldsymbol{I}, \boldsymbol{p})) \tag{5.10}$$

Please see Figure 5.7 for a visualization of the feature extraction for two example annotations.

5.5.4 Displacement Expert Training

In previous steps, we assembled a list of feature vectors with corresponding induced model parameter errors. In this step, the displacement expert $\hat{g}_i^\ell(\boldsymbol{h}_i(\boldsymbol{I}, \boldsymbol{p}))$ is trained on that list to map feature values to the required parameter update as shown in Equation 5.11. Please note that in contrast to g_i^ℓ, the displacement expert $\hat{g}_i^\ell$ is trained on the extracted image features rather than the image data. Therefore, the selection of image features has influence on $\hat{g}_i^\ell$, but not on g_i^ℓ.

$$\begin{aligned} \Delta p_i &= g_i^\ell(\boldsymbol{I}, p_i) = \hat{g}_i^\ell(\boldsymbol{h}_i(\boldsymbol{I}, \boldsymbol{p})) \\ &= \hat{g}_i^\ell[h_i^1(\boldsymbol{I}, \boldsymbol{p}), h_i^2(\boldsymbol{I}, \boldsymbol{p}), ..., h_i^H(\boldsymbol{I}, \boldsymbol{p})] \end{aligned} \tag{5.11}$$

Since we have generated training data and corresponding training data labels, this is a traditional machine learning task and theoretically any regression algorithm is applicable. However, due to the large amount of training features, we

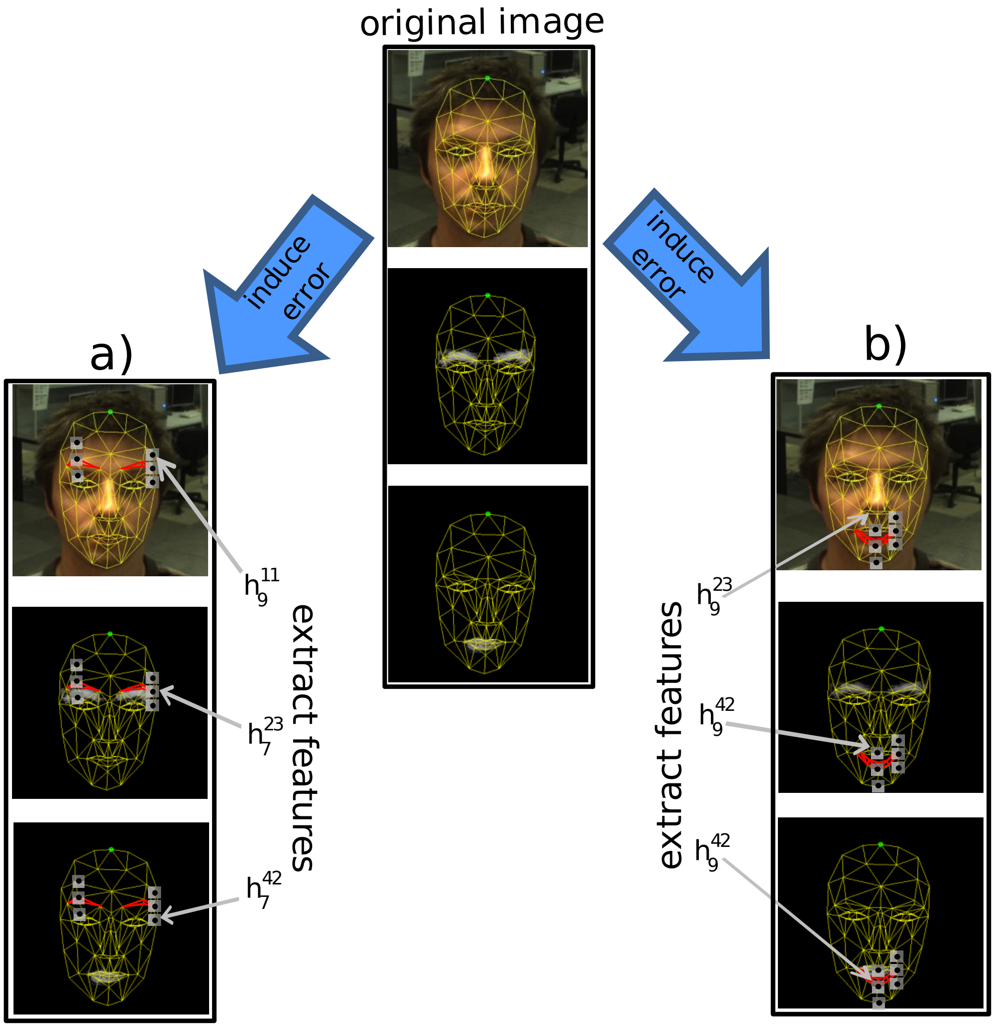


Figure 5.7: Features are extracted at all points influenced by changing a single parameter. Additional features are extracted along the direction of point motion from all image bands. Due to space limitations, only selected image bands are presented here. Top: image bands of the original image with manually fit model. a) error induced in the eye brow raiser parameter. b) error induced in the jaw drop parameter.

choose to integrate model trees to create a mapping of these feature values to the model parameter error [140]. Model trees are similar to well-known decision trees, but replace the class label in the leaf nodes by linear functions. Their strength lies in their capability to select the most relevant features from a large set of training features, therefore no additional feature selection is required. Furthermore, they are very fast at execution time. Equation 5.12 formulates the calculation of the parameter update $\Delta\boldsymbol{p}$ for a previously unseen image. Please note again, that H depends on the parameter.

$$\begin{aligned}\Delta\boldsymbol{p} = \boldsymbol{g}(\mathcal{I}, \boldsymbol{p}) &= (\Delta p_1, \Delta p_2, ..., \Delta p_N)\\ &= (\hat{g}_1^{\ell}[h_1^1(\boldsymbol{I}, \boldsymbol{p}), h_1^2(\boldsymbol{I}, \boldsymbol{p}), ..., h_1^H(\boldsymbol{I}, \boldsymbol{p})],\\ &\quad\ \hat{g}_2^{\ell}[h_2^1(\boldsymbol{I}, \boldsymbol{p}), h_2^2(\boldsymbol{I}, \boldsymbol{p}), ..., h_2^H(\boldsymbol{I}, \boldsymbol{p})],\\ &\quad\ ...\\ &\quad\ \hat{g}_N^{\ell}[h_N^1(\boldsymbol{I}, \boldsymbol{p}), h_N^2(\boldsymbol{I}, \boldsymbol{p}), ..., h_N^H(\boldsymbol{I}, \boldsymbol{p})]).\end{aligned} \tag{5.12}$$

5.6 Application To A Different Model

Trained displacement experts are applicable not only to the Candide-III face model, but also to other shape-based face models. As an example, we demonstrate training a displacement experts to fit an ASM in this section, please see Figure 5.8 for a visualization of the model. The common fitting strategy for ASMs is to search for the best hypothesis $\hat{\boldsymbol{x}}_n$ of each model point $\boldsymbol{x}_n$ individually. Our approach calculates a position update $\Delta\boldsymbol{x}_n = \boldsymbol{g}_n(\boldsymbol{I}, \boldsymbol{x}_n)$ from a displacement expert, again splitting the global displacement expert to local displacement experts. However, these local displacement experts will consider model point positions rather than model parameters. The model parameters are calculated afterwards by minimizing the Euclidean distances between the model points of the projected model and the model points estimated by the displacement experts.

The best hypothesis of each point is computed by $\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n + \Delta\boldsymbol{x}_n$. Again, the core idea is that an ideal displacement expert for fitting a single model point should always determine the correct position update, i.e. it should exactly ”know” the distance to the optimal location, similar to the parameter update in Section 5.5. Therefore, the optimal local displacement experts simply compute the difference between the correct location $\boldsymbol{x}_n^*$ of the n^{th} model point and the current position $\boldsymbol{x}_n$ of the model in the image, see Equation 5.13. If we restrict the choice of $\boldsymbol{x}_n$, for

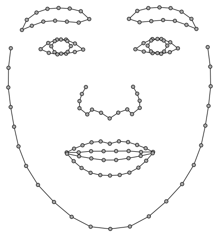


Figure 5.8: We apply our fitting to a ASM as an alternative demonstration.

instance to be located on a straight line or curve, the difference is a single scalar.

$$\boldsymbol{g}_n^*(\mathcal{I}, \boldsymbol{x}_n) = \boldsymbol{x}_n - \boldsymbol{x}_n^* \tag{5.13}$$

Again, the first step is to manually annotate images with the preferred model parameters $\boldsymbol{p}^*$ to specify the preferred model points $\boldsymbol{x}_n^*$. Then, each model point is moved along a line perpendicular to the model contour. The real-valued displacement at each of the positions on this line is computed by $\Delta\boldsymbol{x}_n = \boldsymbol{x}_n - \boldsymbol{x}_n^*$. These known displacements yield the training data for the regression step. We compute image features around the displaced model point. Again, we use Haar-like features of varying size, style and orientation. Finally, for each of the N model points, a regression from image features to known displacements is trained with support vector regression. We chose to integrate a different regression algorithm to demonstrate that our approach is not dependent on the choice of the regression algorithm. However, this choice influences the time required during execution and training. Furthermore, support vector regression requires more memory storage than tree-based induction during training, which restricts its applicability if the number of features is large and not enough memory storage is available.

5.7 Common Refinements

A number of strategies have been proposed with both, objective functions and displacement experts, to increase the fitting accuracy by compensating for local noise. With displacement experts, several measurements are combined to obtain a joint parameter update. With objective functions, optimization refers to the realization of the $argmin$ operator. In this section we briefly introduce the most important strategies for both, objective functions and displacement experts. Note that all strategies ”collapse” to the canonical one if n=1, with n referring to the total number of function evaluations.

5.7.1 Common Refinements on Displacement Expert

As local inaccuracies in Equation 5.1 can lead to large errors, the common idea is to obtain a more stable estimation by combining several measurements. Some common strategies to improve the estimate of $\Delta\boldsymbol{p}$ are depicted in Figure 5.9. In Section 5.8.5 we evaluate the effects of these common refinements to the displacement expert approach.

Canonical

In contrast to objective functions, displacement experts provide a straightforward canonical application: To use one evaluation of $\boldsymbol{g}$ to compute $\Delta\boldsymbol{p}$. This is the direct application of Equation 5.1. Although it is not an refinement, it is mentioned here for the sake of completeness. Please note again, that if perfectly accurate displacement experts could be acquired, the canonical version would always lead to the desired result ($\boldsymbol{p}^*_{\mathcal{I}}$).

Sequential

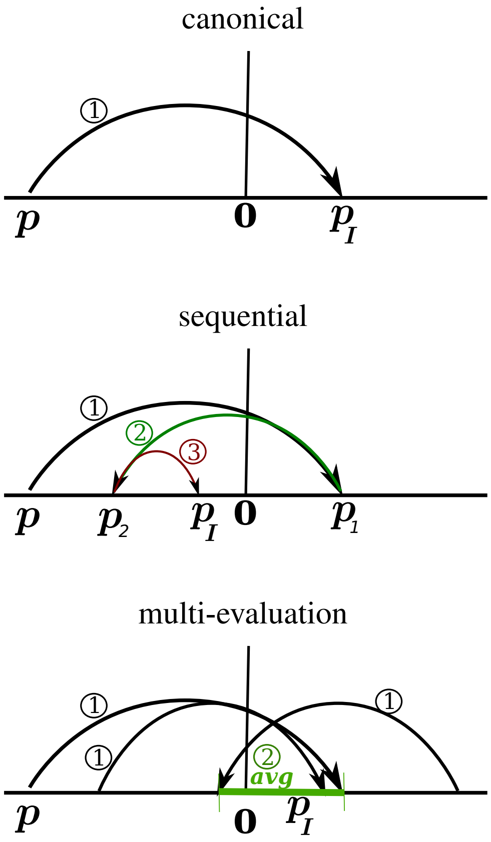
This approach updates the parameters n times sequentially: Step 1: $\boldsymbol{p}_1 = \boldsymbol{p} + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p})$, Step 2: $\Delta\boldsymbol{p}_2 = \boldsymbol{p}_1 + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p}_1)$, $\ldots$, Step n: $\boldsymbol{p}_{\mathcal{I}} = \boldsymbol{p}_{(n-1)} + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p}_{(n-1)})$. The assumption here is that the model parameters "drift" towards the preferred model parameters with every iteration, i.e. that $|\boldsymbol{p}^*_{\mathcal{I}} - (\boldsymbol{p} + \Delta\boldsymbol{p})| \leq |\boldsymbol{p}^*_{\mathcal{I}} - \boldsymbol{p}|$. This also implies that the displacement expert is more accurate near the preferred model parameters. Several example applications have been proposed in literature [61, 116, 77, 145].

Multi-evaluation

This approach computes several $\Delta\boldsymbol{p}$ for n different initial parameterizations $\boldsymbol{p}$. More specifically, it creates additional initial model parameterizations $\boldsymbol{p}_1, ..., \boldsymbol{p}_n$, by sampling around $\boldsymbol{p}$. Then, it calculates the fitted model parameters:

$$\boldsymbol{p}_{\mathcal{I}} = \frac{[\boldsymbol{p} + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p})] + [\boldsymbol{p}_1 + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p}_1)] + ... + [\boldsymbol{p}_n + \boldsymbol{g}(\boldsymbol{I}, \boldsymbol{p}_n)]}{n+1} \tag{5.14}$$

The assumption in this refinement is that the estimation of $\Delta\boldsymbol{p}$ is prone to inaccuracies due to local noise. By combining several estimations, this local noise is compensated for and the combined estimation is more accurate. This approach is used in Particle Filters, such as [131] or [1].



canonical

sequential

multi-evaluation

Figure 5.9: Displacement expert refinement strategies evalute the displacement expert multiple times to ward against local noise. In these examples, $n = 3$ except for canonical.

5.7.2 Common Refinements on Objective Function

The term ”refinements” in the context of objective functions is a bit misleading, since it does not focus on how the function itself is evaluated or applied, but its intent is reducing the number of required evaluations to determine the optimum. As a general rule, search strategies that are computationally more efficient have the disadvantage that they get frequently stuck in local optima. We consider two extreme representatives here to reflect both ends of the spectrum.

Exhaustive Search

This strategy applies the objective function to a densely sampled set of examples and determines the global optimum to fit the model. The advantage is that this strategy is robust against local optima and will determine the global optimum even if the objective function is prone to local noise. On the other hand, if the global optimum is not located correctly or if there are multiple strong optima, this approach is unable to decide, which to chose. Therefore, a small image patch might induce large errors in the fitting.

Gradient Descent

This strategy is applied iteratively to determine the optimum, and is based on numerically calculating the objective function gradient in each step to determine the location of the nearest optimum. The advantage of this approach is that it will determine the optimum must faster than the exhaustive search strategy, specifically, if the initial guess is near the optimum. On the other hand, it is not guaranteed to determine the global optimum but might get stuck in a local optimum.

5.8 Experimental Evaluation

This section presents an evaluation of our approach, which inspects the accuracy of trained displacement experts in several scenarios. We annotate images taken from the ”Labeled Faces in the Wild” database [54] with the ideal model parameters $\boldsymbol{p}^*$ to serve as test data. Due to the size of the database, only images of persons starting with the letter ”A” are considered as a representative subset of 446 images. These images have not been taken with a computer vision application in mind and include many challenges that have to be faced in real-world conditions. Since we capture the training images ourselves, as mentioned in Section 5.5.1, training images and test images are taken from two different data sets, which prevents the displacement experts from specializing on data set properties. To measure the accuracy of the trained displacement experts, we create erroneous model parameterizations $\boldsymbol{p}^{error} = \boldsymbol{p}^{*} + \Delta\boldsymbol{p}_{\mathcal{I}}^{error}$ by inducing errors $\Delta\boldsymbol{p}_{\mathcal{I}}^{error}$ in the manually specified model parameters $\boldsymbol{p}^{*}$. Then, we apply the fitting algorithm to compute model parameters $\boldsymbol{p}_{\mathcal{I}} = \boldsymbol{p}^{error} + \boldsymbol{g}^{\ell}(\boldsymbol{I}, \boldsymbol{p}^{error})$. The distance $|\boldsymbol{p}^{*} - \boldsymbol{p}_{\mathcal{I}}|$ between the preferred parameters $\boldsymbol{p}^{*}$ and the suggested parameters $\boldsymbol{p}_{\mathcal{I}}$ serves as a measurement of accuracy. We denote the remaining error after fitting of a parameter i in the k^{th} image with an artificially induced error of d by $\bar{p}_{i,k,d}$.

A similar approach is taken to evaluate objective function based fitting with the gradient descent strategy. We induce errors in the model parameters in the same way, but then determine the nearest objective function optimum. The model parameters that correspond to this optimum are chosen as $\boldsymbol{p}_{\mathcal{I}}$. However, this approach is not feasible to evaluate the exhaustive search strategy, since their result is independent of the induced error. Therefore, no error is induced in this evaluation and the search algorithm is applied directly to determine the objective function optimum and $\boldsymbol{p}_{\mathcal{I}}$.

Please note that we do not determine the model pose with learned displacement experts in the same manner as we determine the other model parameters. Instead, we first fit an ASM to the image to determine the face contour and positions of eyes and mouth, and calculate the face pose from these reference points.

5.8.1 Evaluation of Common Refinements on Displacement Experts

This section conducts an evaluation of both refinement strategies for displacement experts. The fitting accuracy is visualized in a chart, where the horizontal axis represents the induced error $\Delta\boldsymbol{p}_{\mathcal{I}}^{error}$ and the vertical axis represents the average remaining error $|\boldsymbol{p}^{*} - \boldsymbol{p}_{\mathcal{I}}|$ in all images after fitting. Therefore, a single measurement in these charts refers to the average of $\bar{p}_{i,k,d}$ for a fixed i and fixed d.

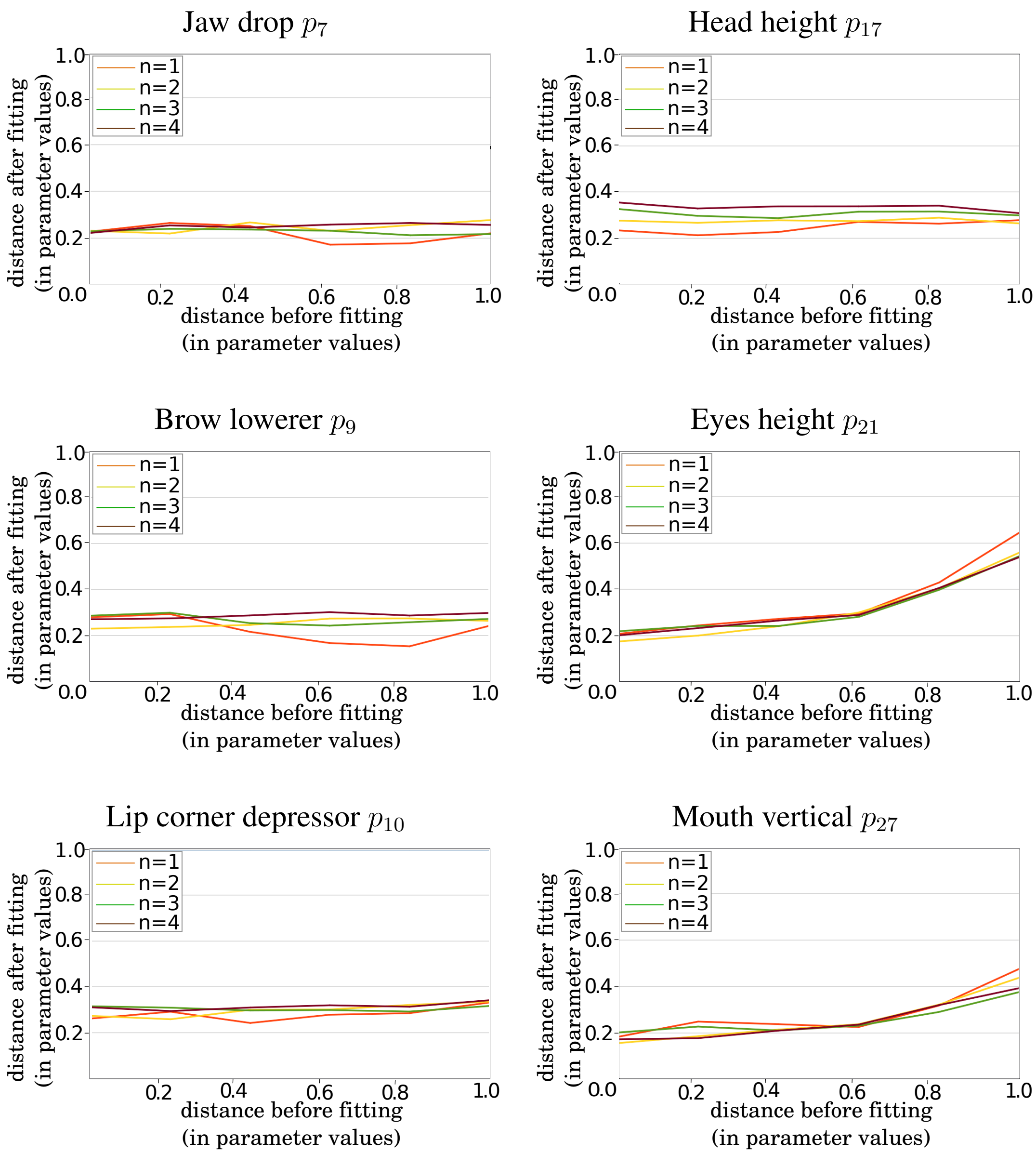
For the sake of readability, we present only a subset of the model parameters here, since our evaluation creates one chart per parameter. The parameters are chosen to reflect the whole face area and neither stress nor neglect certain face regions, providing a representative subset of all parameters. Furthermore, some of the model parameters reflect changes that have been proven difficult to annotate in the test images, such as parameter p_{24}(Cheeks Z extension), parameter p_{14}(Lid tightener) or parameter p_{15}(Nose wrinkler). Therefore, the evaluation would be conducted against erroneous training or test data annotations and would not reflect the fitting approach accuracy.

Figure 5.10 demonstrates that most parameters do not gain much from iterated application of the fitting. The two notable exceptions are parameters p_{20} and p_{27}. Even in these cases, the accuracy is not improved further after the second iteration. The reason is that the fitting accuracy is not influenced by the induced error to a large extend. Even when the induced error is small, there is a constant remaining error after fitting, which contradicts the underlying assumption of this refinement strategy. Therefore, the first iteration reduces the fitting error beneath this threshold and further applications do not increase the accuracy any more. Once the error is within this threshold region, the local displacement expert either suggests slightly wrong updates or no update at all. A further observation is that with parameter p_{17} each additional iteration reduces the accuracy rather than increasing it. This is caused by a small number of images, in which the estimation of the induced error fails due to covered facial components, and the local displacement expert suggests random parameter updates. This causes the induced error to increase instead of decrease. With each additional iteration, these errors increase ever more. The same effect is observable with parameters p_{20} and p_{27}. According to the chart of parameter p_{20}, the error is expected to be 0.64 after one iteration, when the initial error is 1.0. Fitting with an initial error of 0.64 is expected to result in an error of 0.3 in turn. Therefore, the fitting result after 2 iterations with an initial error of 1.0 is expected to result in an error of 0.3. However, this is not the case. Again, the reason is that some images induce highly erroneous fitting results.

In contrast, the multi-evaluation strategy proves much more beneficial, as the charts in Figure 5.11 demonstrate. There is a significant gain in accuracy between $n = 1$ and $n = 2$ for all example parameters. The accuracy is increased further with any additional multi-evaluation. However, the benefit decreases with every additional application. Furthermore, already with $n = 2$ even large initially induced errors are compensated, especially for parameters p_{20} and p_{27}.

5.8.2 Evaluation of The Derived Objective Function

This section provides an evaluation of fitting the model with an objective function that is obtained from a local displacement expert, as demonstrated in Section 5.4. There is a large number of optimization algorithms available to the community, but the most accurate result is always obtained by exhaustive search, when applicable. Since in our application the search space is limited by the semantic


Jaw drop p_7

Head height p_{17}

Brow lowerer p_9

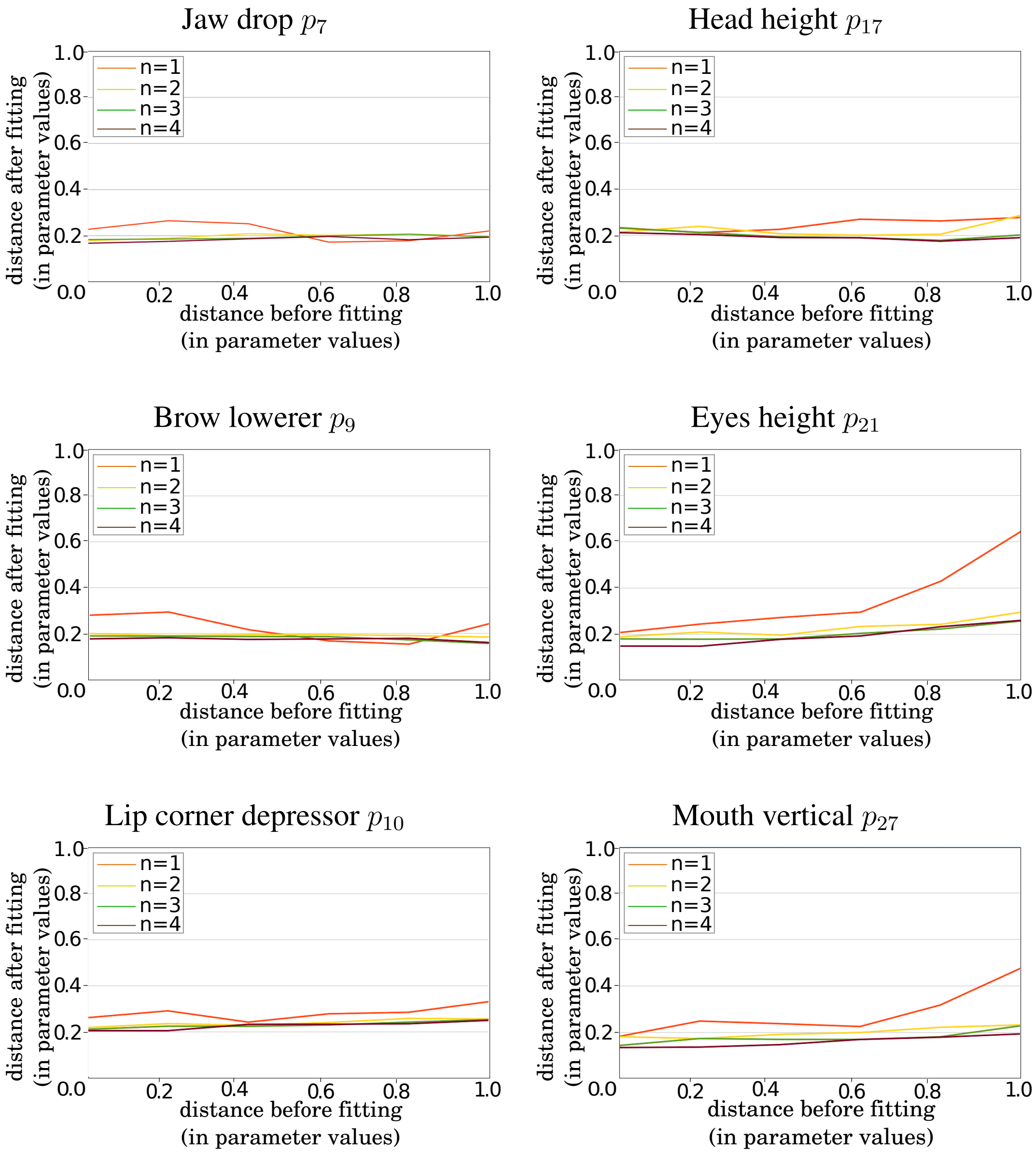
Eyes height p_{21}

Lip corner depressor p_{10}

Mouth vertical p_{27}

Figure 5.10: The sequential approach does not improve the fitting results very much even after several iterations.

interpretation of the model parameters, we apply heuristic search to determine the optimal objective function value. Please note that this approach is not dependent on the initially induced error. Since this approach is very time-consuming we



Jaw drop p_7

Head height p_{17}

Brow lowerer p_9

Eyes height p_{21}

Lip corner depressor p_{10}

Mouth vertical p_{27}

Figure 5.11: The multi-evaluation fitting strategy improves the fitting accuracy significantly with each step.

integrate gradient descent as a faster optimization algorithm for comparison.

In this evaluation, we do not present single charts for single parameters, but summarize the results for the gradient descent strategy in a single chart in Fig-

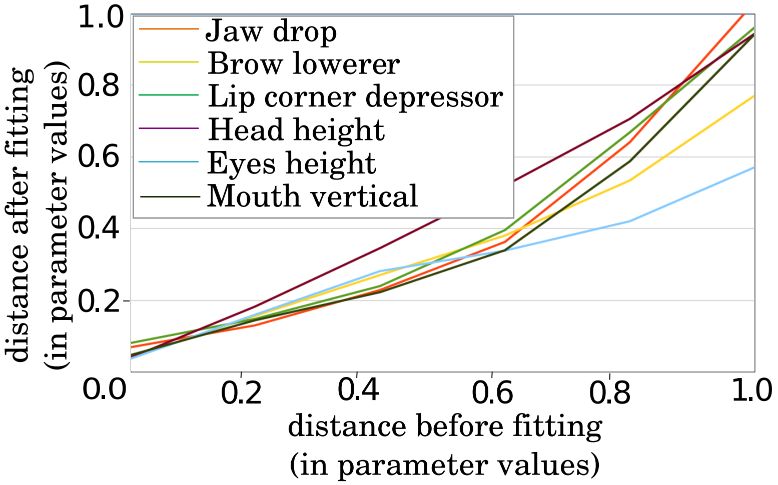


Figure 5.12: Objective Functions with gradient descent for selected parameters.

ure 5.12. This approach is very accurate when the induced error is small. However, this is rather the result of local noise since the optimization is caught in a local optimum after only a few iterations. Results for the exhaustive search are provided in Table 5.1. This strategy results in an accuracy that is comparable to or slightly larger than the threshold value of the displacement expert approach.

5.8.3 Absolute Fitting Accuracy

To compare the fitting accuracy of different model parameters, we chose a different visualization that is independent of the induced error. Our evaluation measures the fraction of models that have at most a specific error in a specific parameter after fitting, i.e. we visualize the cumulative error distribution of $|\boldsymbol{p}^* - \boldsymbol{p}_{\mathcal{I}}|$ in all images as shown in Figure 5.13. Therefore, this evaluation considers all $\bar{p}_{i,k,d}$ with i being fixed. For instance, 70% of all models are fitted with an error of 0.2 or less in the ”Jaw drop” parameter. The local displacement experts are provided the complete set of image bands and a multi-evaluation with $n = 4$ is applied in this evaluation. Induced parameter errors range from -1.0 to 1.0. Again, we present only a subset of the complete parameter vector. Parameters are chosen that are especially important for facial expression recognition.

We observe that the local displacement experts significantly compensate the induced error. However, there is a strong gap between the two parameters “Eye lid raiser“ and ”Lip stretcher” and the other facial expression related parameter. The

Jaw drop p_7	0.28
Lip corner depressor p_9	0.24
Brow raiser p_{10}	0.29
Head height p_{17}	0.36
Eyes height p_{20}	0.31
Mouth vertical p_{21}	0.23

Table 5.1: The fitting accuracy of objective function with exhaustive search strategy is independent of the initial error.

reason is that these parameters cause the smallest variations in the face and are difficult to observe due to the image size. Therefore, estimating the correct value is difficult, not only for the computer but also for the human annotator. Please note that the automatic fitting is limited by the accuracy of the manual annotations in the training data and the evaluation is limited by the manual annotation of the test data. Another observation is that, in a few cases, the initial error is even increased by the fitting algorithm. This is represented by the models with errors larger than 1.0. These errors occur because of heavy occlusions of facial components due to glasses or large beards that cause the calculation of image bands to fail, and cause the fitting approach to behave unpredictably. However, again, human annotators are also unable to specify the exact location of these facial components. Figure 5.14 depicts some example images with covered facial components that occur in the test data.

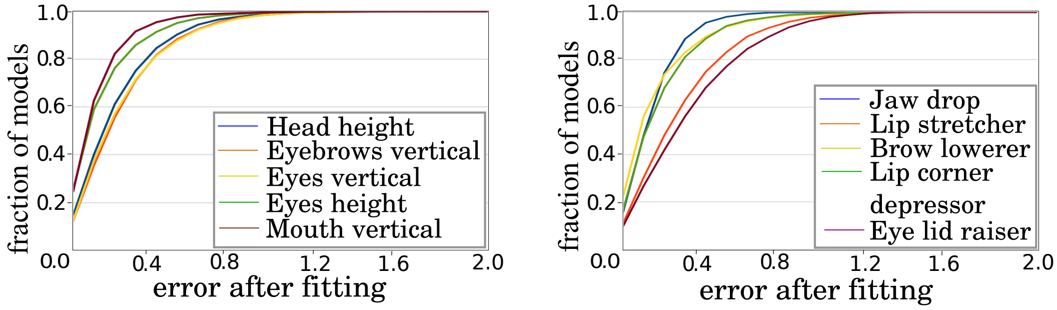


Figure 5.13: These graphs visualize the cumulative error after model fitting. Initial errors induced range from -1.0 to 1.0 with eleven evaluations per test image.

5.8.4 Impact of Provided Image Bands

In this experiment, the impact of provided image bands on the fitting accuracy is inspected. A displacement expert that is trained on the original image data only serves as a fixed reference baseline for comparison. Then, displacement experts are trained from the same training data utilizing additional image bands. Please note that displacement experts are applied to the same image representation that they have been trained on. To determine the accuracy of a single displacement expert, the experimental setup described in Section 5.8.3 is used. However, instead of creating a cumulative error distribution, the error is computed as the average of all single parameter error values of all parameters utilized in Figure 5.13 and for all test images. Therefore, this evaluation calculates the average of all $\bar{p}_{i,k,d}$ with $i \in \{7, 8, 9, 10, 16, 17, 18, 19, 21, 27\}$.

Errors in Table 5.2 are given with respect to the baseline displacement expert. The first row shows, which image band has been added to train the displacement expert. Image bands are added subsequently and single displacement experts are represented by table columns. Since every displacement expert is provided one additional image band, the number of image bands provided increases with the table column index. For instance, the displacement expert represented by the third column has been provided $\mathcal{I}, I_{skin}, I_{lip}$, and its error is 80.7% of the baseline error. Please note that not all local displacement experts rely on all available image bands. Instead, only those image bands that add valuable data for the local displacement expert are integrated.

Since error values are reduced with increasing table column index, providing additional image bands increases the fitting accuracy. The largest impact is observed when I_{lip} is added to the image representation. The reason is that many parameters benefit from this information, since many parameter describe the mouth

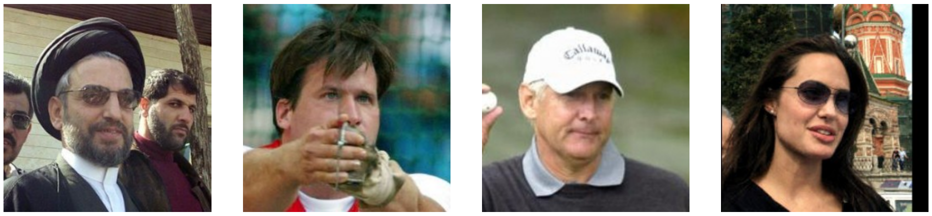


Figure 5.14: Covered facial components, for instance due to glasses or beards, cause the computation of image bands in the multi-band image representation to fail.

image bands	$\mathcal{I}$	+I_{skin}	+I_{lip}	+I_{brow}	+I_{retina}
error rate	100.0%	92.5%	80.7%	79.0%	78.5%

Table 5.2: Providing additional image bands reduces the fitting error. One image band is added per column.

position and shape. Theoretically, this information is also present in I_{skin}, since the mouth results in a black area within the face region. However, inspecting Table 4.7 shows that classification of skin also has the smallest accuracy compared to the other image bands and therefore contains a large amount of noise. As a result, when other image bands are added to the image representation, they are chosen by the model tree training since they are less noisy, which increases the fitting accuracy.

A second experiment inspects the benefit of specific image bands on single parameters. The assumption is that a model parameter referring to the shape of the lips does not benefit from I_{brow} as much as from I_{lip}. To verify this assumption, Table 5.3 provides a comparison of four displacement experts on different combinations of image-bands. Most local displacement experts gain from providing I_{skin}. However, the image bands I_{lip} and I_{brow} are beneficial only to those displacement experts that model corresponding facial components. For instance, the $Jaw\ drop$ parameter gains heavily from I_{lip}, but not from I_{brow}. This is reasonable, because I_{brow} does not provide any valuable information to determine the degree of opening of the mouth.

parameter	$\mathcal{I}$	$\mathcal{I}, I_{skin}$	$\mathcal{I}, I_{skin}, I_{lip}$	$\mathcal{I}, I_{skin}, I_{brow}$
jaw drop	100.0%	81.2%	**46.7%**	84.1%
lip stretcher	100.0%	72.1%	**62.3%**	72.9%
lip corner depressor	100.0%	100.0%	**86.6%**	100.0%
brow lowerer	100.0%	93.4%	95.8%	**90.4%**
head height	100.0%	100.0%	**78.2%**	**74.4%**

Table 5.3: Local displacement experts gain more from image bands that refer to the modeled facial component than from other image bands.

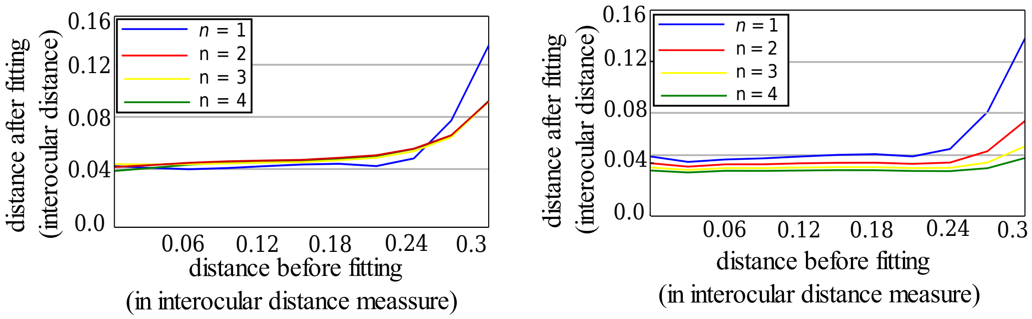


Figure 5.15: The two different refinement strategies for displacement experts. Left: iteration. Right: multi-evaluation.

5.8.5 Evaluation On A Different Model

This section demonstrates that integrating multi-band images is beneficial to fitting other shape-based models, as well. We change the face model in this experiment and integrate an ASM [19].

Test images are again taken from the ”Labeled Faces in the Wild” database and errors are artificially induced into manually specified model annotations. However, to follow the common evaluation strategy for these models, errors are induced by displacing model points directly instead of inducing errors in the model parameters. Then the fitting algorithm is applied and we again calculate a cumulative histogram of errors. However, in this evaluation, the error is not measured in parameter differences, since this is not an intuitive error measure with this model, but in model point distances. We calculate the average distance of the fitted model points to the corresponding manually annotated model points. To compensate for varying face size in the images, all distances are computed in interocular distance measure, which refers to the distance between the pupils in the image. To inspect the fitting accuracy, again a cumulative error histogram is calculated.

Refinements on Displacement Experts

This evaluation determines the impact of the refinements presented in Section 5.7.1 on the fitting accuracy. In Figure 5.15, the average error after fitting is plotted against the initially induced error, similar to Figures 5.10 and 5.11. Each final displacement is an average taken over all local displacements and test images, i.e. the global fitting error of a model is computed to be the average of all local fitting errors.

From these graphs, we draw the following conclusions: First of all, performing several iterations of the sequential approach only has a significant effect if the initial displacement is large ($\Delta \boldsymbol{x}_n$ >0.2), and this effect levels off after n=2 iterations. This finding is consistent with the observations presented in Section 5.8.1. To inspect the reason for this, we calculate the standard deviation of the errors after fitting for $n = 1$ in Figure 5.16. Although it decreases with the induced error, it is always larger than 0, which leads to the conclusion that there is always some noise in the estimation, even if the error is small. If the error is below this threshold, iterated applications of the displacement experts merely suggests parameter updates that cause an oscillation around $\boldsymbol{x}_n^*$. This explains why further iterations have no significant effect, as the second iteration already often ends up in this area.

Clearly, multi-evaluation benefits the most from increasing n, again. The reason is that the different evaluations of the displacement expert are independent of each other, and with more samples at different locations, the noise in the estimation of $\Delta \boldsymbol{x}_n$ is reduced as is indicated by the findings presented in Figure 5.16.

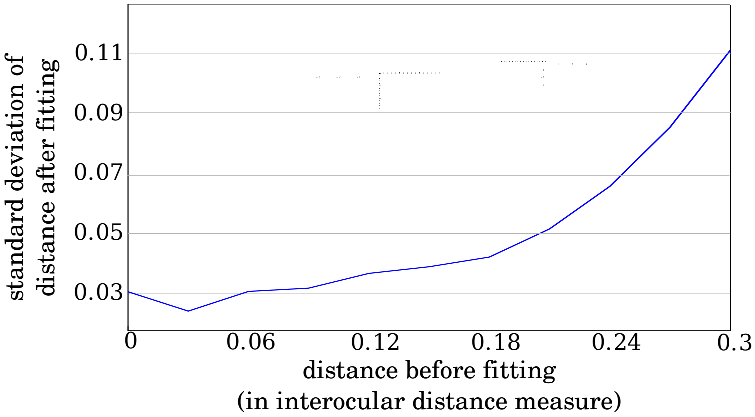


Figure 5.16: The standard deviation of the displacement expert predictions increases with larger initial displacements.

Refinements on Objective Functions

As depicted in Figure 5.17, the gradient descent strategy provides high accuracy for small values of $\Delta \boldsymbol{x}_n$. Exhaustive search always determines the global minimum and therefore represents the best possible result that can be achieved with an objective function. These findings are similar to those for applying objective function to fitting the Candide-III face model in Section 5.8.2.

Impact of Provided Image Bands

This experiment demonstrates that not only fitting the Candide-III face model, but also the ASM benefits from providing additional image bands. As can be seen in Table 5.4, including additional image bands increases the accuracy with every image band added. In Table 5.4, again one image band is added per column. The largest impact is observed when I_{skin} is added. However, adding additional image bands again increases the accuracy by approximately 7%. Therefore, not only fitting the Candide-III face model but also the ASM fitting benefits from integrating multi-band images.

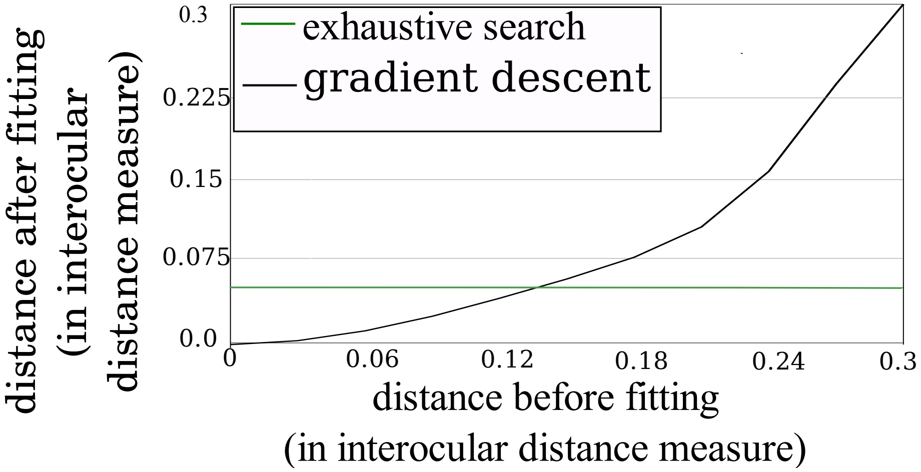


Figure 5.17: Comparison of gradient descent and exhaustive search in fitting with objective functions.

image bands	$\mathcal{I}$	$+I_{skin}$	$+I_{lip}$	$+I_{brow}$	$+I_{retina}$
error rate	100.0%	91.4%	89.0%	86.7%	84.9%

Table 5.4: Providing additional image bands reduces the fitting error also for the ASM. Again, one image band per table column is added.

5.9 Discussion

We presented an approach to fit the Candide-III face model to previously unseen images. Multi-band images and trained displacement experts are applied for this task. Our fitting strategy has been evaluated on the ”Labeled Faces in the Wild” database, which contains images in unstructured environments. Furthermore, we directly compare model fitting with objective functions and with displacement experts on two representatives.

Our approach offers two benefits: Firstly, the Candide-III model is independent of training data, which is not the case with face models like ASMs and AAMs. The Candide-III model is hand-designed and therefore its parameters provide semantic information, without the need for further interpretation. In contrast, the parameters of ASMs and AAMs refer to variations in manually annotated training data and to not necessarily provide a semantic interpretation. This renders it specifically applicable in scenarios where no information about the person in front of the camera is available or a large number of different persons have to be considered. An example for such a situation will be presented in Chapter 7, where an experiment with a large number of participants is conducted and no previous information on the participants is available for training. Another imaginary example would be an application in a public place, like a store window.

Secondly, in direct comparison, displacement experts are clearly superior to objective functions. They are more robust to local noise, due to the multi-evaluation refinement, and are much faster, since they are evaluated less often. However, although our fitting strategy has been developed with the Candidie-III face model in mind, it is not restricted to this face model. With shape-based models, annotating training images is straightforward. Furthermore, there is no restriction towards the modeled object, but if other objects than faces are considered the consequence is that new classifiers for multi-band image have to be created.

The drawback of the Candide-III face model is that it is not objectively guaranteed that it reflects human face shapes or face motions well. This assumption is based on the believe in the designer’s experience. Furthermore, the same drawback that has been mentioned for ASMs and AAMs, is true also for trained displacement experts. Care must be taken, that the images annotated for training offer a wide variety of different faces. If the training images do not reflect the variety of human faces well, for instance if they are trained only on male persons or no bearded men are included, the algorithm will not be able to work well if confronted with random images. We fell for this trap when taking images to train displacement experts in the evaluation. Although we took care to take images with changing lighting, background, gender, complexion, facial hair, facial expression and head pose, the images did not include covered facial components. The integration of multi-band images compensates this effect to some extend, but places the same requirement on the images that are used to train the pixel-based classifiers.

Chapter 6

Facial Expression Recognition

Already Charles Darwin in the nineteenth century conducted scientific research on facial expressions and published his book *The Expression of the Emotions in Man and Animals* in 1871. He already noted commonalities in facial expression between humans, independent of age and ethnicity, stating that "the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements" [25, p. 352]. However, Darwin himself was inspired in his research by previous work, conducted by French neurologist Guillaume Duchenne and his book *Mcanisme de la Physionomie Humaine*. He described mimetic muscles and stimulated muscle contractions by electrical current, artificially generating facial expressions, such as depicted in Figure 6.1 [30]. Duchenne also distinguished smiles, which do not include activity of the muscle orbiting the eye from those that do. These genuine smiles have later been termed *Duchenne smiles* in his honor by Ekman et al. [33].

This research lead to the idea of facial action units, atomic facial activities formed by combinations of muscle contractions. Although, the idea was first introduced by Hjrts [51], it became popular with the work of Ekman and Friesen. Their Facial Action Coding System (FACS) is the most comprehensive explanatory system for facial activity devised so far [32]. Action units describe combinations of muscle contractions and combinations of action units describe facial expressions. The FACS has a total of 32 action units involving facial muscles, and another 23 action descriptors, which include actions like turning and tilting of the head that do not directly involve facial muscles. Table 9.1 in the Appendix comprehensively lists the 32 action units and their corresponding muscles, based on the 2002 edition of FACS [35]. Action units may vary in intensity, with the intensity score given as a Latin letter ranging from A (least discernible intensity) to E (maximum intensity). For example, a notation of AU 1E would signify a brow raised to maximum possible intensity. Every individual has its own specific maximum intensity for every action unit, therefore these intensities cannot be directly translated into metric measurements.

To link between facial expressions and emotions, Ekman and Friesen describe six universal facial expressions - happiness, sadness, anger, disgust, fear and surprise - that are expressed and interpreted in the same way by humans of any origin all over the world, independently of ethnic or cultural background, and that correspond to specific emotions [31]. Examples for these universal facial expressions, taken from a standard database, are depicted in Figure 6.2. Precise definitions for the facial action inherent to these universal facial expressions are given by the same authors in the *Emotional FACS* (EmFACS) [39].

Although these facial expressions themselves are universal, their intensity, i.e.

the degree of muscular facial activity, may vary between individuals, dependent on cultural background and personal emotional baselines. For instance, the facial expression regarding disgust is generally performed with greater intensity for people with Japanese cultural background when compared to people with U.S. American cultural background, while the other universal facial expressions are performed comparatively less intense [80].

In the late sixties, psychological interest in non-verbal communication led to a number of publications concerning facial expressions. Studies conducted by Mehrabian indicate that facial expression is the major factor in face-to-face verbal

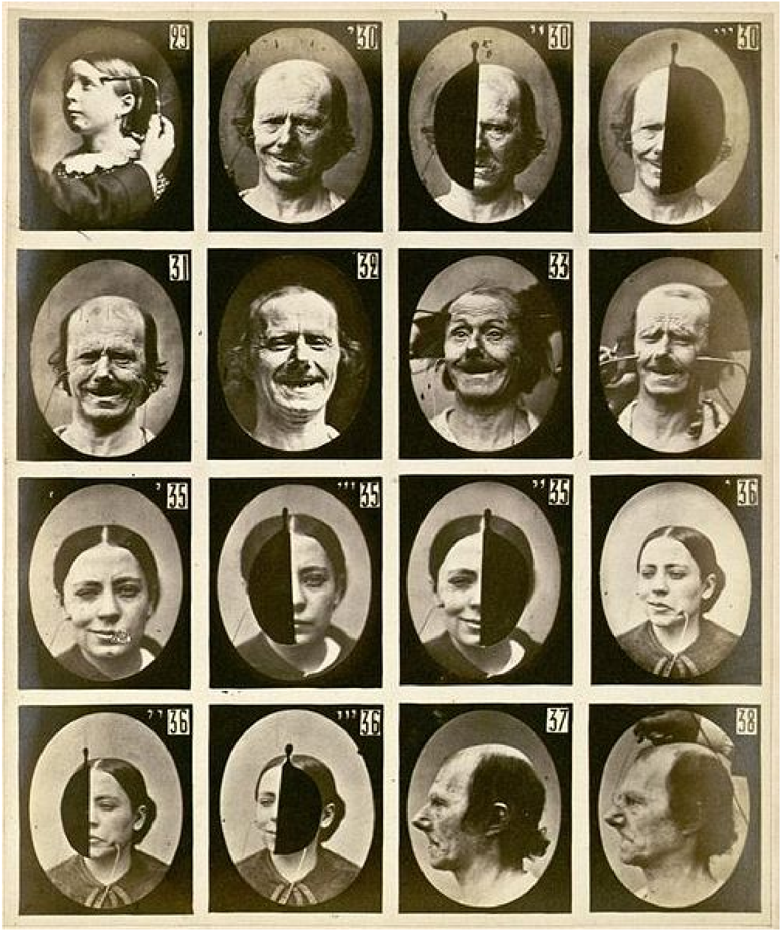
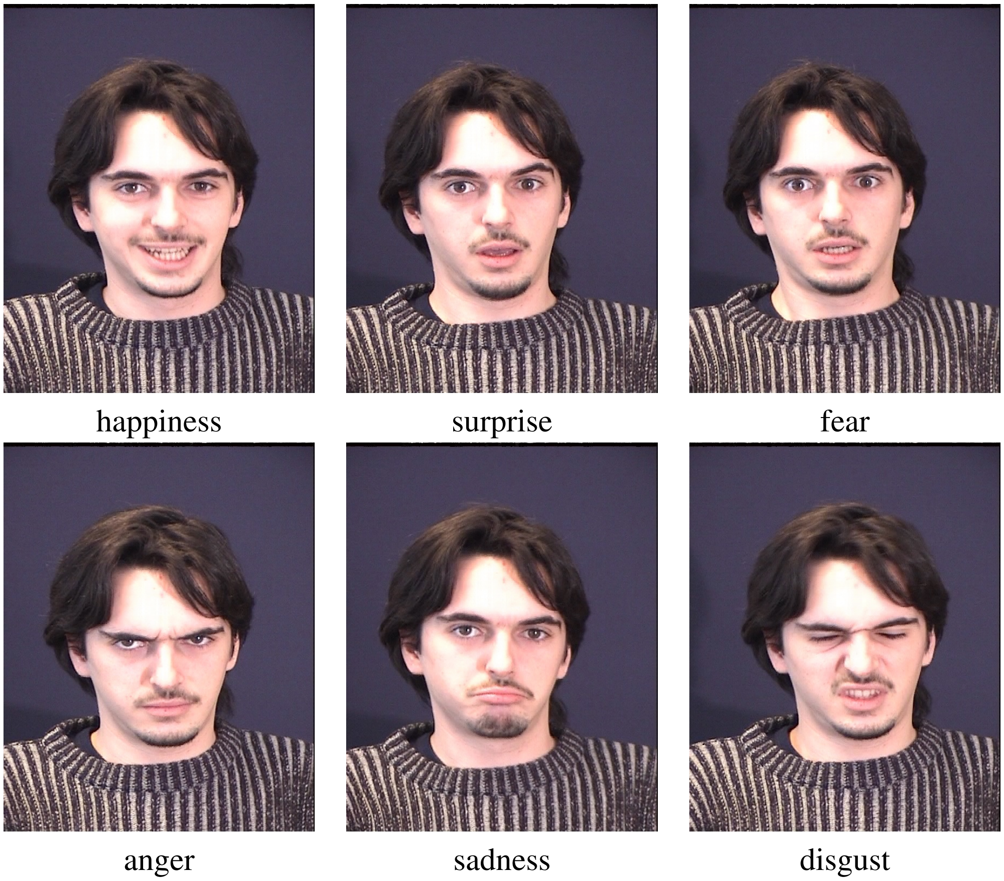


Figure 6.1: Facial expression created artificially by electrical stimulation of the facial muscles, taken from [30, p. 277].



happiness surprise fear

anger sadness disgust

Figure 6.2: Example images of the six universal facial expressions, taken from the MMI database.

communication, conveying 55 percent of the total message effect, with intonation (38 percent) and wording (7 percent) contributing the remainder [85].

In nature, facial expressions occur as a process, developing from the neutral face or a previously shown expression into the facial expression apex, i.e. the moment of greatest expression intensity. The process of facial expression therefore consists of three phases: onset, apex, and offset [97]. With regard to visual data, humans are able to recognize facial expressions from still images as well as video sequences, indicating that for human facial expression recognition temporal aspects of facial expressions are redundant.

In this chapter, we apply a model-based approach to automatically recognize facial expressions from single images. We determine model parameters for single images to obtain a numeric representation of the face state. Model parameters are gained by fitting the Candide-III face model as described in Chapter 5 and manually correcting failed fitting.

6.1 Problem Statement

Traditionally, facial expression algorithms are trained and evaluated on a set of standard databases to test their performance. Mostly, these databases are structured to depict examples of the universal facial expressions or activation of single action units. Since it is well-known that the accuracy measured on training data is not representative for the classifier quality, most approaches are evaluated with stratified cross-validation or leave-one-out-validation [42]. Such techniques, which use different subsets of a database for training and testing, are referred to as *self-classification evaluation* techniques. Since the universal facial expressions are, as their name implies, universal, the choice of the actual training and test database should not have any influence on the classifier quality and classifiers should generalize well even across databases. Unfortunately, this is not the case.

Following this idea, we draw the conclusion that good performance of an algorithm on a specific test-set does not guarantee that this algorithm will perform well also on other test sets (or real-world applications). The reason is that misleading information might be included in the database itself due to design decisions that were taken during the database assembly. This has been confirmed in several research fields in- and outside of computer vision, raising doubt on using self-classification evaluation as the only measure of quality. However, this is still the predominant strategy for empirical evaluations in computer vision.

A well-known challenge when applying any machine-learning technique is to avoid overfitting. This describes a process, in which the classifier focuses on memorizing the training data rather than determining the underlying function. In doing so, the classifier looses its capability to generalize on data not explicitly provided in the training data set [104]. Comparing self-classification evaluation and cross-database evaluation can also be considered as observing a kind of overfitting on a more complex level. In this case the classifier does not memorize the training data itself but properties of the training database and therefore structures in the data that were not intended by the database author. Cross-database evaluation is an alternative to self-classification evaluation, and obtains training and test data from different databases, instead of sampling them from the same database.

6.2 Solution Idea

Instead of aiming for high recognition results in self-classification evaluation, which is prone to lead to specialized classifiers that are tuned towards single databases, we instead inspect and evaluate classifiers for facial expression recognition across databases. In this approach one database is used for training and another database is used for testing. This procedure helps spotting classifiers with superior generalization properties and separates them from algorithms, which achieve high accuracies for self-classification evaluation because of their ability to learn special characteristics of the training database. Our goal is to determine, which classifier leads to the best generalization over multiple databases and to balance between recognition accuracy and generalization. Therefore, when confronted with a choice between alternative approach, instead of selecting the one with the highest accuracy in self-classification evaluation, we compare them in cross-database evaluation. In doing so, we take an important step towards obtaining a robust facial expression classification algorithm with high generalization properties.

6.3 Related Work

Benchmarking and comparison is an important part of (computer vision) research. This fact is contributed to by publishing surveys, comparisons of different algorithms and databases. For instance, Neilon et al. compare algorithms for correspondence matching in stereo images [92]. They find that evaluation in this area is usually based on determining error rates on example image pairs and assuming the algorithm with higher accuracy to be superior. However, inspection of the statistical significance shows that this measurement might be misleading and might not reflect the correct accuracy ranking of the algorithms. Similar conclusions are drawn by research groups in the area of image segmentation, where quality measurement is a difficult task due to the fact that everybody has his/her own demands on the result and no general measurement exists [154, 41]. Most approaches work well on specific groups of images only. Therefore, Zhang et al. propose to combine basic evaluation techniques with machine learning to generate a domain independent framework. Similar to the recognition of facial expressions, image segmentation or automatic image retrieval is often a subjective manner and defining objective quality measurements is difficult [125]. Go et al. tackle this by relying on the labeling of multiple persons during their approach.

For the purpose of face image analysis, a number of publicly available databases have been proposed [96, 153, 45, 87, 63, 138]. An overview of the databases that are important for this thesis will be presented in Section 6.4. Mostly, evaluation of (facial expression recognition) algorithms is conducted on these databases in self-classification evaluation. Example facial expression recognition approaches that are evaluated in this manner are: [67, 150, 153, 100]. Several research groups extend this evaluation strategy by conducting self-classification evaluation on multiple databases [148, 155, 5]. However, each database is inspected separately. For more details on the methodologies utilized in these approaches and the technical details, we refer to Section 3.2.

We also took this approach in our earlier work, and presented facial expression recognition from a live camera stream and from database images [83, 127]. The approach taken is similar to the one presented in this thesis, as it uses the Candide-III face model and is trained on the complete database instead of neutral and apex images only. However, evaluation is conducted only on the CK database and MMI database and is very tuned towards these databases, since stratified cross-validation is performed on image level. In contrast, we evaluate our approach in cross-database evaluation in this thesis, which serves two purposes: Firstly, to obtain more representative results, secondly, to create classifiers that generalize well. In contrast to our earlier work, cross-database evaluation is a major step of improvement.

6.3.1 Cross-database Evaluation

Some research groups have already further extended the idea of including multiple separate databases by conducting cross-database evaluation. The conclusion drawn from such experiments is that getting good results in self-classification evaluation is not indicative of achieving a good performance in general. An early example outside computer vision is the work of Livshin et al [75]. They conducted an experiment with five established sound databases, and find that accuracy drops from 98% in self-classification evaluation to only 20% when tested on another database. Similar results have been demonstrated for emotion recognition in speech [120].

A very recent example for this insight is given by Whitehill et al. who state a warning from relying too much on self-classification on established databases. They state that ”It is conceivable that by evaluating performance on these data sets the field of automatic expression recognition could be driving itself into algorithmic ’local maxima’. ” [55]. They gained this insight by applying the algorithm they presented in earlier work [74] in real-world environment. Since they achieved high accuracy in self-classification evaluation, they assumed robustness in real-world, as well. However, their recent work concludes that this expectation was misleading, although they pointed out already in their earlier work that results in cross-database evaluation were not as promising as in self-classification evaluation. The same observation is confirmed by Shan et al. [121]. Again, high accuracies are reported in self-classification evaluation as well as a sever loss of recognition rate in cross-database setups. Similar examples are presented by Koelstra et al. [114, 66]. Although the difference between self-classification evaluation and cross-database evaluation is not so severe in their work, it is still notable.

6.3.2 Conclusion

Several research groups have demonstrated that results gained in self-classification evaluation are not representative for the accuracy obtained in a more unstructured setup [114, 55, 121, 120]. However, most approaches are evaluated in self-classification evaluation, sometimes mentioning reduced results in cross-database evaluation without paying much attention to this effect. In contrast, we propose to evaluate algorithms in cross-database evaluation rather than self-classification evaluation, since this is a better indicator of the generalization capabilities of a classifier. We conclude that although 98% is an impressive result, a classifier that scores 80% across databases would be preferable.

Some research groups present recognition results that are significantly higher than the ones presented in this work, however, these results are mostly tuned towards a single database, which raises doubt on their capability to generalize. Kotsia et al. present high recognition rates on the CK database, but their feature extraction is heavily tuned towards this database, since they manually specify small sets of FACS action units that occur only with a single expression in the database [67]. Hong et al. report very high recognition results on the CK database and MMI database when they include neutral images in the recognition [52]. However, since they do not mention any image weighting or selection criterion of images, it is very unclear on which subsets of the database their algorithm has been trained and evaluated. Martin et al. present high recognition results on the FEEDTUM database with an AAM-based approach [79]. However, the AAM has been generated from the very images that are used for training and evaluating the classifiers, which tunes their approach towards this image data.

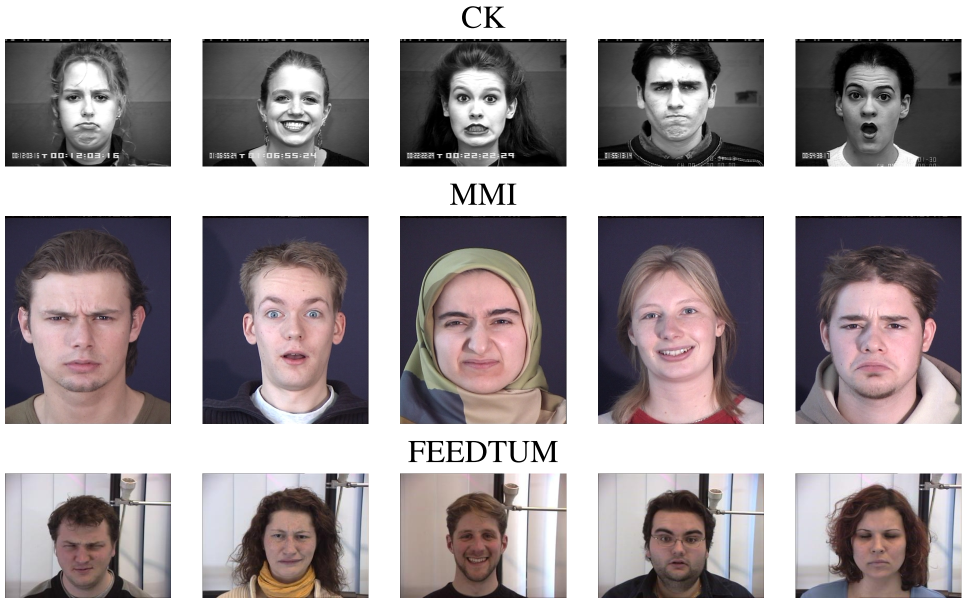
6.4 Databases

Several databases have been proposed containing face images for computer vision purposes. These databases consist either of single images or image sequences, mainly captured in lab or office environments. Mostly, several images or image sequences per subject are available. Usually, information on the database content is provided by the database authors. Depending on the database purpose, this so-called ”meta-data” includes person identity, facial expression, action unit activation, annotated facial components or landmarks.

The database content is selected with a specific application in mind. Facial expressions are a known challenge in face identification and therefore databases intended for face identification often also depict facial expressions. They offer a large number of different subjects. However, those databases usually consist of single images rather than image sequences and the facial expression depicted is not necessarily given in the meta-data. Therefore, these databases mostly do not provide good data to train facial expression recognition systems. In contrast, databases intended for facial expression recognition provide single images or image sequences that are labeled with one of the universal facial expressions or FACS action unit activations. Usually, this labeling is provided on a per-sequence basis. Figure 6.3 depicts some example images from the three databases that have been used in this thesis.

6.4.1 Cohn-Kanade Database

The first applied database is the *Cohn-Kanade Facial Expression Database* (CK database), consisting of 2105 video sequences taken from 182 adult subjects, both male and female and of varying complexion[63]. All video sequences were shot from frontal views with uniform lighting conditions and little or no out-of-plane movement. The sequences are digitized into a format of 640×480 pixels and range from 6 to 90 frames in length. The subjects were asked to perform various facial actions, including the six universal facial expressions as described by Ekman et al. The video sequences have been annotated by experts according to the EmFACS. This database is the most often referred to collection of image data for facial expression recognition.



CK

MMI

FEEDTUM

Figure 6.3: Some randomly chosen images from each of the three databases.

6.4.2 MMI Facial Expression Database

The *MMI Facial Expression Database* (MMI) contains more than 184 sequences of images showing facial expressions or single action units, taken from a total of 19 subjects [96]. It contains both frontal and profile-view sequences of facial expressions, as well as some sequences depicting both frontal and profile view simultaneously. In contrast to the image sequences in the CK database, these sequences do not solely show facial expressions developing from neutral faces to maximum intensity, but also show the subsequent phase of expressions declining from maximum intensity to neutral faces, with sequence length varying from 40 to 520 frames. In our evaluation, we use a subset of the MMI database consisting of 108 sequences. The selection criterion is that only frontal-view images are depicted. Concerning metadata, expert annotations are provided on a per-sequence basis, similar to the annotations provided for the CK database.

6.4.3 FEEDTUM Database

The $FEEDTUM$ database was captured in 2006 at the Technische Universität München and contains image sequences depicting the six universal facial expressions as they have been defined by Ekman et al [138]. Furthermore, example image sequences that depict neutral faces are included, however they were not considered in this thesis, since the other databases do not provide neutral image sequences. The image data is captured from a total of 18 subjects, each displaying each of the seven facial expressions (six universal facial expressions and a neutral face) three times. Instead of asking the participants to completely act the facial expressions, short movie clips and still images were used to induce emotions and provoke more natural facial expressions.

6.5 Data Annotation

Because the databases provide their meta-information on a per-sequence basis, but we will train image-based classifiers, facial expression labels for single images have to be specifically derived. We use the construct of facial expression expressiveness to determine this label.

6.5.1 Automatically Labeling Single Images

In the CK database, the final frame is always the apex. Since the index of the apex frame is not given in the MMI and FEEDTUM database, we annotate it manually. The class label $\mathcal{C}_{\mathcal{S}} \in \{surprise, happiness, sadness, fear, anger, disgust\}$ is the annotation for the entire image sequence $\mathcal{S}$. To acquire class labels for single images we apply a heuristic: We define the expressiveness $\mathcal{E}$ of a frame as the linear interpolation between the closest apex frame (with $\mathcal{E}$=1) and neutral frame (with $\mathcal{E}$=0). For instance, if frame I_i is neutral, and frame I_{i+n} is the closest apex, then an intermediate frame I_k with $i < k < (i+n)$ has an expressiveness of $\mathcal{E}_k = (k-i)/n$. The class label $\mathcal{C}_k$ is determined by applying a threshold:

$$\mathcal{C}_k = \begin{cases} \mathcal{C}_{\mathcal{S}}, & \text{if } \mathcal{E}_k \geq \theta \\ neutral, & \text{if } \mathcal{E}_k < \theta. \end{cases} \tag{6.1}$$

Since the point of transition from neutral face to facial expression is not well-defined, choosing the value of θ is left to the designer. For the classifiers presented in this thesis we apply a threshold of $\theta = 0.33$. Some example images with an expressiveness of approximately 0.33 are presented in Figure 6.4. For every image, we extract a feature vector that is labeled with the image's expression class label. We refer to this labeled feature vector as a training or test ”observation”. However, the feature extraction itself will be detailed later in Section 6.6.1.

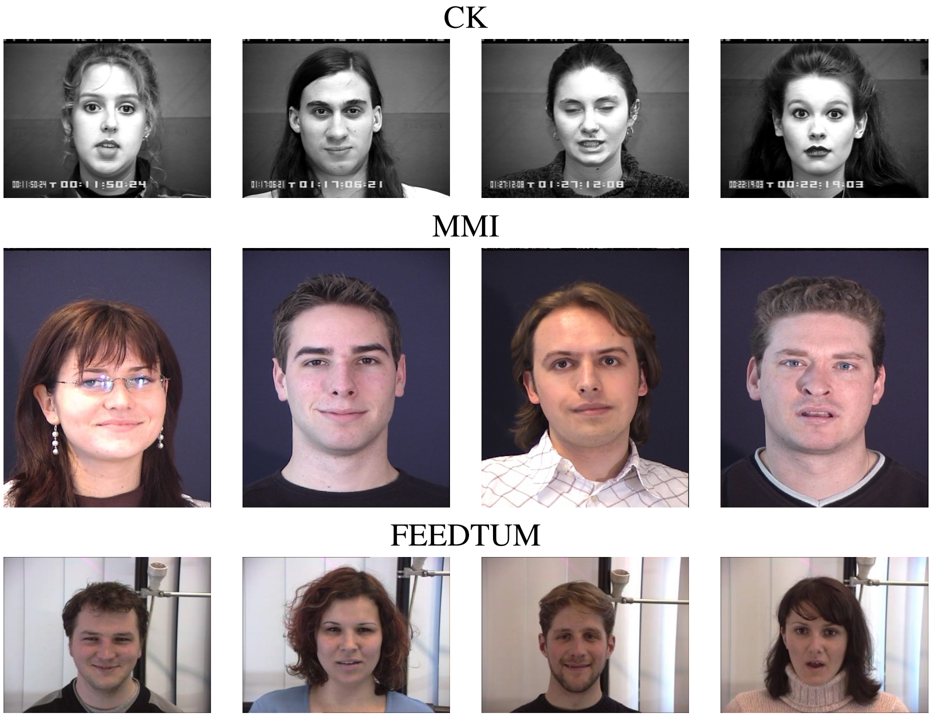


Figure 6.4: The point of transition marks the image in an image sequence when a neutral face changes to a displayed facial expression. These images depict facial expressions with an expressiveness of $\theta = 0.33$.

6.5.2 Selecting Observation Weights

If there are different costs for misclassifying observations, different weights can be associated with each observation. Usually, the machine learning algorithm treats every observation in the same manner, not preferring any observation over another or trying to maximize its parameters on a subset of observations only. In this case, all observations have the same, constant weight.

$$W_k^{constant} = 1.0. \tag{6.2}$$

However, more complex weighting schemes are applicable as well. For instance, if only neutral and apex frames are of interest, we are able to remove the remainder from training and classification by adjusting their weights to 0. The constant weight is granted to neutral and apex images, only. The advantage of this weighting scheme is that the neutral and apex frames have expert-based labels that are provided with the meta-data of the database. On the other hand, a large amount of available data is not used in this approach.

$$W_k^{extremes} = \begin{cases} 1.0, & \text{if } \mathcal{E}_k = 1.0 \\ 1.0, & \text{if } \mathcal{E}_k = 0.0 \\ 0.0, & else \end{cases} . \tag{6.3}$$

This leads to the idea of weighting images according to the clarity of facial expression display. Since the facial expression is most difficult to determine at the point of transition between the neutral face and the facial expression, images near this point in the image sequence get smaller weights than images near the neutral or apex frame. Intermediate images are weighted by linearly interpolating between these images in the image sequence. Please note that the second line is important only if decreasing expressiveness is depicted, too. We visualize the different weighting Functions in Figure 6.5.

$$W_k^{linear} = \begin{cases} \mathcal{E}_k, & \text{if } \mathcal{E}_k \geq \theta \\ 1 - \mathcal{E}_k \cdot \frac{1-\theta}{\theta}, & \text{if } \mathcal{E}_k < \theta \end{cases} . \tag{6.4}$$

6.6 Facial Expression Recognition

In this section, we describe the procedure to recognize facial expression from database images. We fit the face model to every image of the databases and extract the face model parameters to obtain a single observation. Then, we train a classifier on a collection of these observations. Please see Section 5.4 for more information on the face model.

6.6.1 Feature Extraction

The only features relevant to facial expression classification are the facial expression parameters $\boldsymbol{a}$ of the face model. Therefore, we extract these features directly from the model parameters as a subset of $\boldsymbol{p}_\mathcal{I}$ and call them the feature set $\boldsymbol{p}^{FPS}$. The features $\boldsymbol{p}^{FPS}$ are person-specific. For example, the degree, to which a person presses the lips together in a neutral face differs among individuals. To acquire more person-independent features $\boldsymbol{p}^{FPI}$, we compute the difference between the features $\boldsymbol{p}_k^{FPS}$ of the k^{th} image in the sequence, and those of the first image in the sequence $\boldsymbol{p}_1^{FPS}$, which is given to be neutral: $\boldsymbol{p}_k^{FPI} = \boldsymbol{p}_k^{FPS} - \boldsymbol{p}_1^{FPS}$ The union of both feature sets is $\boldsymbol{p}_k^{FPP} = \boldsymbol{p}_k^{FPS} \cup \boldsymbol{p}_k^{FPI}$. Please note again, that features are extracted for all images of each database and that every image is inspected individually for the extraction of the features $\boldsymbol{p}^{FPS}$.

6.6.2 Classification

We utilize Support Vector Machines (SVMs) with radial basis function kernels for classification. The optimal parameters are determined by an extensive search over possible parameter settings.

Since SVMs calculate a single separating hyperplane, they are not directly applicable to problems with more than two classes. However several extensions have been proposed to apply them also to multi-class problems. A very straightforward

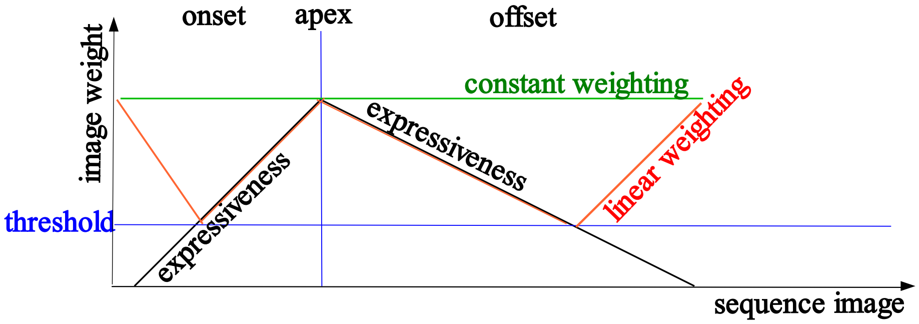


Figure 6.5: Different observation weights are computed from the image expressiveness.

idea is to train several SVMs, each designed to distinguish between two classes only. A test observation is presented to all SVMs and voting on the classification results determines the final class assignment. We rely on a further improvement of this idea, which was proposed by Hastie [48]. Instead of applying a simple voting algorithm to the classification results, a joint probability distribution of all classes is constructed from the single SVM results and the class label with the highest probability is selected.

6.7 Experimental Evaluation

This section presents an evaluation of the facial expression classification and inspects different weighting schemes and feature sets. We use three databases, three feature sets and three observation weighting schemes. Since the databases differ in size a lot, we extract two randomly chosen subsets of each database, one to serve as training and one as test data. The random samples are chosen that way that images in the training and test split are not taken from the same image sequences. Furthermore, the size of the sampled subsets does not depend on the database size but is the same for all databases. Accuracy values for classifiers are calculated by dividing the summed weights of all correctly classified observations in the test split by the summed weights of all observations in the test split, see Equation 6.5. In this equation, $\mathcal{C}_\mathcal{O}$ refers to the class label of the observation $\mathcal{O}$, $W_\mathcal{O}$ refers to its weight and $\hat{\mathcal{C}}_\mathcal{O}$ refers to the classifier prediction of the observation. The results presented are obtained by calculating the average of five randomized splits. We conduct a chi-square test of significance with a significance level $\alpha = 0.05$ on all experiment including classification to determine the impact of feature set selection and weighting scheme selection. Results that are not significantly different are marked in the following way: Both results are marked with ^{*x}, where x is a number that references non-significant pairings.

$$r = \frac{\sum\limits_{\hat{\mathcal{C}}_\mathcal{O}=\mathcal{C}_\mathcal{O}} W_\mathcal{O}}{\sum W_\mathcal{O}}. \tag{6.5}$$

6.7.1 Traditional Approach

A common way to determine class labels for single images from the sequence label is to ignore non-apex images. Therefore, this experiment focuses on apex expressions only, and ignores all other images completely. Please note that the disadvantage of this approach is that any facial expression is interpreted to be one of the universal facial expressions, since there is no neutral class. The fact that there is a significant gap in accuracy between the CK database, the MMI database and the FEEDTUM database already hints to differences in the database structure, an observation experienced by other research groups as well [148]. The CK and the MMI databases are assembled by persons that are very familiar with the FACS, the apex images in these databases are in general depicting stronger facial expressions and in a more structured manner than in the FEEDTUM database. Furthermore, the fraction of data used from a single database differs. Since two images per sequence are selected, the ratio of sequence length and number of sequences determines the fraction of data rejected. Since the CK database offers a large number of short sequences, the fraction of data accepted from this database is highest.

The results presented in Table 6.1 are comparable to results published by various researches earlier, mostly published on the CK and MMI database [150, 13, 122, 4, 114, 148, 5], which demonstrates that our approach is within state-of-the-art techniques

6.7.2 Cross-database Comparison

According to the theory of universal facial expressions, they are depicted similar in all databases, since they are independent of age or culture. Therefore, a classifier trained on any of the databases should determine facial expressions independent of context conditions, even across databases.

Unfortunately, as the results presented in Table 6.2 demonstrate, this is not the case. The reason is that the database data is biased by expectations of the database designer and influenced by design decisions, like sequence length, number of participants or participant instructions. Classifiers trained from such biased data rely on database properties, preventing them from generalizing across databases. Accuracy values are significantly higher when training data and test data is taken from the same database than when they are taken from different databases. We will apply cross-database evaluation in the remainder of this chapter and refer to a combination of training and test databases as ”scenario”.

Another drawback of this approach is that we are not able to model the transition between neutral faces and those depicting a facial expression even if neutral images are included. This decision is taken by the classifier training, based on the provided training data.

CK	MMI	FEEDTUM
84.1%	79.8 %	67.2%

Table 6.1: Approaches that are trained and evaluated on apex images only achieve high accuracy values.

training	test database		
database	CK	MMI	FEEDTUM
CK	84.1 %	60.3 %	33.9%
MMI	66.2%	79.8 %	36.6%
FEEDTUM	56.6%	58.9 %	67.2%

Table 6.2: Accuracy values across databases are significantly lower that in self-classification evaluation.

6.7.3 Evaluation of Weighting Functions

In this experiment, we integrate the neutral class and model facial expression intensity to increase linearly from neutral images to apex images, as presented in Section 6.5.2. We assume all images with $\mathcal{E}_k \leq 0.33$ to depict a neutral face and all other images to depict the facial expression provided in the database meta-data.

This increases the amount of data for training and evaluation by a great extend, since there are no more observations with $W_O = 0$. The factor by which the data is increased depends on the average sequence length of the database's image sequences. It is roughly 15 for the CK database and even more for the other databases. However, this also induces fuzzy image labellings that correspond to images at the point of transition from a neutral faces to facial expressions, which are more likely to be misclassified. Therefore, there are two oppositional effects it terms of accuracy.

training	test database		
database	CK	MMI	FEEDTUM
CK	71.7%	53.3%^{*1}	31.6%^{*2}
MMI	47.5%	60.0 %	41.9%^{*3}
FEEDTUM	39.3%	45.8 %	53.5%

Table 6.3: Taking the complete data into account with constant observation weighting adds more data to the evaluation, but also increases the amount of fuzzy data.

training database	test database		
	CK	MMI	FEEDTUM
CK	76.2%	53.2%^{*1}	32.3%^{*2}
MMI	49.3%	61.6 %	42.7% ^{*3}
FEEDTUM	41.9%	48.7%	55.3%

Table 6.4: Applying the linear weighting scheme balances between amount of training data and data fuzziness.

Please note the ratio of neutral images to images depicting facial expressions. Only one third of each sequence depicts a neutral image, which means that per sequence there is twice as much data for the facial expression then for the neutral face. However, since we have six facial expressions, in total there is three times the data for neutral faces than for depicted facial expressions. To compensate for this imbalance in order not to emphasize neutral faces over facial expressions, the weight of neutral images has to be divided by three. The factor required to balance neutral and non-neutral images is calculated according to Equation 6.6, with a referring to the required value. The idea behind this formula is to balance the integrated observation weights values of neutral images and integrated observation weights of non-neutral images, while considering that there are six different facial expressions.

$$a \cdot 6 \cdot \int\limits_{\mathcal{C}_k = neutral} W_k = \int\limits_{\mathcal{C}_k \neq neutral} W_k \tag{6.6}$$

We apply Equation 6.6:

$$a \cdot 6 \cdot \frac{1}{3} = \frac{2}{3} \rightarrow a = \frac{1}{3}$$

As Table 6.3 indicates, in general the accuracy drops, which demonstrates that the effect of fuzzy training data is stronger than the increase in training data quantity. Please note that despite the loss in accuracy, this approach still holds the advantage of hand-modeling the point of transition between neutral and non-neutral faces.

To balance between data quantity and data clarity we propose to apply the linear weighting function of Equation 6.4. The idea is that the nearer an image is to the point of transition, the more fuzzy the facial expression label gets. Therefore, the observation weight is highest for neutral and apex images and decreases towards the point of transition. Table 6.4 demonstrates that all classifiers in any possible scenario gain from this weighting. Although some of these increased accuracy values are not significant, the tendency is evident.

Obviously, more complex weighting functions would be applicable as well to introduce expert knowledge on facial expression intensity modeling. However, to avoid subjectivity in the data, we restrain from using such complex weighting models.

6.7.4 Evaluation of Feature sets

As mention in Section 6.6.1, we propose a set of features consisting of two subsets. This evaluation inspects the impact of the feature set selection on the classification accuracy. Again, the evaluation is conducted on all different database scenarios. Tables 6.5, 6.6 and 6.7 present accuracy values for classifiers that are trained with different feature sets. Classifiers using $\boldsymbol{p}^{FPS}$ perform worst in almost any scenario. Adding feature set $\boldsymbol{p}^{FPI}$, which results in feature set $\boldsymbol{p}^{FPP}$, greatly improves the classification accuracy, as a comparison of Table 6.5 and Table 6.7 reveals.

However, providing only feature set $\boldsymbol{p}^{FPI}$ increases the accuracy in most scenarios, again. This demonstrates that providing more data is not ensured to increase classification accuracy, since a classifier might focus on misleading data. In this case, the feature set $\boldsymbol{p}^{FPP}$ includes also the features $\boldsymbol{p}^{FPS}$, which contain information specific for the depicted person. Therefore, if the database consists of images taken from only a small number of subjects, the classifier will rely on person-specific information. In contrast, the $\boldsymbol{p}^{FPI}$ feature set considers facial motion as it is induced by facial expressions rather than facial structure. Since accuracy values are lowest in any database combination for $\boldsymbol{p}^{FPS}$, we consider these features to be very weak.

6.7.5 Specialization properties

All classifiers evaluated in a self-classification evaluation perform comparably well when trained on $\boldsymbol{p}^{FPI}$ or $\boldsymbol{p}^{FPP}$ features, therefore we take a closer look at the cross-database performance. Since the aim is to train classifiers that are not specialized on a single database, accuracy values across databases should be near the value in the self-classification scenario of the target database. The idea behind this approach is that self-classification evaluation represents the highest accuracy value that is obtainable on a database. A second motivation is that a

training	test database		
database	CK	MMI	FEEDTUM
CK	61.4%	45.2 %	27.5%
MMI	41.3%	60.9%^{*4}	36.6%
FEEDTUM	34.2%	44.6 %	51.7%^{*6}

Table 6.5: Evaluation across databases with linear weighting scheme and feature set $\boldsymbol{p}^{FPS}$ only.

training	test database		
database	CK	MMI	FEEDTUM
CK	78.9%	53.2%^{*5}	36.5%
MMI	60.8%	61.3%^{*4}	46.6%
FEEDTUM	52.3%	51.4%	51.6%^{*6}

Table 6.6: Evaluation across databases with linear weighting scheme and feature set $\boldsymbol{p}^{FPI}$ only.

training	test database		
database	CK	MMI	FEEDTUM
CK	76.2%	53.2 ^{*5}%	32.3%
MMI	49.3%	61.6^{*4} %	42.7%
FEEDTUM	41.9%	48.7%	55.3%

Table 6.7: Evaluation across databases with linear weighting scheme and feature set $\boldsymbol{p}^{FPP}$, repeated from Table 6.4 for simpler comparison.

classifier, which generalizes perfectly would ”hide” on which database it has been trained. Therefore, we calculate a measurement for the specialization of a classifier as the mean difference of its accuracy in cross-database evaluation and the self-classification evaluation value of the target database. The intention is to spot classifiers that are specialized on their training database and that as a consequence do not generalize well. The higher the value is, the more a classifier is specialized to a single databases. Specialization is calculated according to Equation 6.7, where $r^F_{d_1 \rightarrow d_2}$ refers to the accuracy of a classifier that is trained on database d_1 and evaluated on d_2 using feature set F. Note, that if $d_1 = d_2$, it represents a self-classification evaluation setup. The databases $d_1 \in \mathcal{D}$ and $d_2 \in \mathcal{D}$ are taken from $\mathcal{D}$, which refers to the set of databases $\mathcal{D} = \{CK, MMI, FEEDTUM\}$. Table 6.8 presents specialization properties for all scenarios and feature sets. The mean generalization value is 13.8 for the $\boldsymbol{p}^{FPI}$ feature set and 19.7 for the $\boldsymbol{p}^{FPP}$ feature set, which indicates that the $\boldsymbol{p}^{FPI}$ feature set is more robust. As mentioned already in Section 6.7.4 this strengthens our finding that more data does not necessarily improve the robustness of the classifier. Furthermore, there is a gap in specialization between the CK and MMI database and the FEEDTUM database. This indicates again the similarity of the CK and MMI database.

$$g^F_d = \sum_{\hat{d} \in \mathcal{D} \text{ } without \text{ } \{d\}} \frac{r^F_{\hat{d} \rightarrow \hat{d}} - r^F_{d \rightarrow \hat{d}}}{|\mathcal{D}| - 1} \tag{6.7}$$

In a follow-up experiment, we train classifiers on two databases and evaluate them on the third, remaining database. The $\boldsymbol{p}^{FPI}$ features are used in all scenarios. The idea of this experiment is to observe the effect of adding data in a cross-database setup. As Table 6.9 depicts, classification accuracies are close together. When the MMI or CK database is added for training, classification accuracy is higher than in any corresponding 1-on-1 experiment. For instance, the accuracy $r_{FEEDTUM \rightarrow CK} = 52.3\%$ if only the FEEDTUM database is utilized is increased to $r_{MMI,FEEDTUM \rightarrow CK} = 61.5\%$ when the MMI database is added. The same effect is observable in the second scenario where the MMI database is added: $r_{CK \rightarrow FEEDTUM} = 36.5\%$ increases to $r_{CK,MMI \rightarrow FEEDTUM} = 49.9\%$. As mentioned, this effect is also evident with the CK database: $r_{MMI \rightarrow FEEDTUM} = 46.6\%$ increases to $r_{CK,MMI \rightarrow FEEDTUM} = 49.9\%$ and $r_{FEEDTUM \rightarrow MMI} = 41.9\%$ increases to $r_{CK,FEEDTUM \rightarrow MMI} = 58.7\%$. However, the effect levels off when the FEEDTUM database is added. This is reasonable, since it has a high specialization value, and there is no benefit when adding it to a database with a lower specialization value. However, even the CK database and the MMI database ben-

feature set	training database		
	$\boldsymbol{p}^{FPS}$	$\boldsymbol{p}^{FPI}$	$\boldsymbol{p}^{FPP}$
CK	20.0%	17.6%	21.8%
MMI	11.6%	11.5 %	18.2%
FEEDTUM	15.7%	19.7%	23.6%

Table 6.8: The specialization is significantly lower with $\boldsymbol{p}^{FPI}$ in all scenarios.

	test database		
	CK	MMI	FEEDTUM
result	61.5%	58.7 %	49.9%

Table 6.9: When multiple databases are combined for training, the evaluation results are less widespread.

efit from merging, since they also have a specialization larger than 0.0.

6.8 Regression

In some applications not only the facial expression itself but also the facial expression intensity might be of interest. Therefore, we train regression algorithms that determine the expressiveness from the face model parameters. We train one regressor per facial expression that determines its intensity. Please note that therefore these regressors are provided with information that is not available in the classification approach: the facial expression itself. Furthermore, since the neutral class can be interpreted as depicting any facial expression with a very small intensity, no regressor is trained for the neutral faces.

To evaluate the regressors, we calculate the correlation between the regressor prediction and the expressiveness for all test images. Table 6.10 depicts the average correlation of all facial expressions and in all scenarios. Due to the observation in the previous sections, only the $\boldsymbol{p}^{FPI}$ features are presented here. A striking observation is that regressors obtain high correlation values when tested on the CK database, even higher than in the self classification scenario. This is reasonable, since the highly exaggerated facial expressions in this database support the task of estimating intensity. Please note, again, that the facial expression is known in this setup and merely the intensity is determined. Since, intuitively speaking, the apex frames in this database depict stronger facial expressions than

training	test database		
database	CK	MMI	FEEDTUM
CK	0.83	0.45	0.43
MMI	0.75	0.64	0.42
FEEDTUM	0.70	0.49	0.55

Table 6.10: Evaluation across databases with regression, linear weighting scheme and $\boldsymbol{p}^{FPI}$ features.

in the MMI and FEEDTUM database, the regression algorithm is challenged with the task to "extrapolate" the facial expression intensity over the training data. In contrast, in the classification setup, the classifier is challenged with data that is not reflected very well in the training data.

6.9 Across Facial Expressions

In this section, we inspect the accuracy of classification with respect to single facial expressions. Instead of calculating the results from different database splits, only one split is used but the confusion matrix is calculated from all scenarios. The matrix in Table 6.11 states the probability that a facial expression is confused with any other facial expression. For instance, the probability that a surprised face is classified as a fearful face is 9.7%. As Table 6.11 indicates, the most stable facial expressions are surprise, happiness and neutral.

Most confusions are between neutral faces and facial expressions. This is expected, especially for sadness, since it reflects the cases near the point of transition and this facial expression involves only small changes in the face. However, because it is well distinguishable from other facial expressions, there are only small confusion with other facial expressions. Please note that these confusions are symmetric and there are many images depicting a neutral face that are confused to depict a sad face.

A common confusion, which looks striking at first glance, is the confusion from fear to happiness, since these emotions are quite oppositional. However, they cause similar movements in the face, like stretching the lips. The same holds true for the confusion between disgust and anger, which are mainly distinguishable by either pressing the lips together or having the mouth lightly opened.

We perform a similar experiment to inspect the facial expression intensity estimation. Again we train one regressor per facial expression and calculate the

True	Classified as						
lable	neutral	surprise	fear	happiness	sadness	disgust	anger
neutral	76.2	2.4	4.3	2.8	10.5	1.9	1.9
surprise	4.4	83.2	9.7	0.0	0.9	1.4	0.0
fear	25.6	15.5	31.0	12.6	1.6	11.4	2.0
happiness	9.0	0.0	2.9	74.4	2.3	8.0	2.7
sadness	40.6	2.0	5.8	0.0	44.6	3.3	2.9
disgust	18.1	0.0	4.6	8.8	4.7	43.7	20.6
anger	19.9	1.7	2.1	1.2	17.6	20.3	37.1

Table 6.11: This matrix depicts the probability that a facial expression is classified as another and is calculated from all scenarios.

average correlation in all scenarios in Table 6.12. Happiness and surprise have a high rate of correlation, which is reasonable, because those facial expressions include the most facial movement.

6.10 Discussion

In this chapter, we presented our approach to determine facial expressions from single images. We applied classifiers to determine the class label from model parameters. In contrast to the traditional approach, which considers apex images only, we included the complete database data. The main focus has been the evaluation of the approach, where we decided to use cross-database evaluation instead of traditional self-classification evaluation.

Integrating the complete database in the training and test data holds advantages, but also some disadvantages. One of the advantages is that it reflects real-world conditions better and provides more flexibility in designing what is actually considered a facial expression and what is still a neutral face. Otherwise the maximum-margin classifier will determine the split point ”in the middle” between a neutral face and an apex expression, which depends on the database content. This is an advantage if more subtle facial expressions are of interest. However, the main drawback is that the point of transition from neutral face to facial expression has to be specified, which is clearly a subjective decision. In this thesis, we utilized a simple heuristic for this decision, but we did not prove that this meets the perception of people well. More subjective decisions are induced by the choice of the weighting function. Although, we consider that this thesis took a clear step,

surprise	fear	happiness	sadness	disgust	anger
0.65	0.56	0.75	0.37	0.61	0.51

Table 6.12: evaluation across facial expressions with regression

we are well aware that from this beginning much open research opportunities still exists.

The advantage of the proposed cross-database evaluation is that it provides a comparison of algorithms without the influence of database bias. It prevents algorithms from being tuned towards single databases. This benefit becomes obvious by comparing the different feature sets. If confronted with the task to chose one to integrate in a running system this decision can be made either on the accuracy in self-classification evaluation or cross-database evaluation. When inspecting the results in self-classification evaluation, both feature sets $\boldsymbol{p}^{FPI}$ and $\boldsymbol{p}^{FPP}$ seem to be comparable in accuracy. The superiority of the feature set $\boldsymbol{p}^{FPI}$ is visible only in cross-database evaluation, hinted by our experiment on specialization in section 6.7.5. If guided by accuracy values in self-classification evaluation only, the choice might even be to use the apex-only classifiers introduced in Section 6.7.1. However, inspecting their specialization (26.4 for CK, 24.3 for MMI and 23.7 for FEEDTUM) reveals their drawback. Please note, however, that cross-database evaluation does not provide information how well a database corresponds to real-world data. This decision is still left to the user and we decided this to be true with the databases utilized in this thesis. Therefore, a valuable extension to this approach would be data that is actually captured in real-world conditions. However, if such data is available, evaluation will still benefit from cross-database evaluation. Specialization on the database still prevents classifiers from generalizing well, even in this case.

Chapter 7

Applications

Human-machine interaction traditionally relies on a small number of devices, of which keyboard, mouse, screen and speakers are the most common. However, more advanced communication mechanisms have also been envisioned, inspected and sometimes integrated. Often, these mechanisms are inspired by human-human communication. In other cases, they are specifically tuned towards a single application. An imaginary example is given in the movie ”Minority Report”, where a gesture-based interface is depicted with the idea in mind that this interface allows for a much faster interaction than traditional human-machine interaction. However, recent developments like the Wii-Controller or Microsoft’s Kinect sensor shift such interfaces from imagination close to reality. The Wii-Controller has been developed by Nintendo and is tracked via a infrared tracking bar. It is designed for interactive gaming, but is also quite feasible for gesture recognition [132]. Since the drawback of the Wii-controller is that it has to be held in hands, vision-based approaches for gesture recognition have also been proposed [88]. Very recently, Microsoft published the Kinect sensor which is able to obtain registered optical/depth images and therefore allows for the reconstruction of colored point clouds [142].

A large area, in which advanced interface technologies are applied is movie production. For instance, depicting realistic faces with believable facial expressions is an important topic in the movie industry, with well-known examples being the movies ”Matrix: Reloaded” and ”The curious case of Benjamin Button” [10]. The production of these movies relies on a motion capture technique, which determines the facial expression of an actor to render a face with the same facial expression in the movie scene. Since humans are very trained on recognizing and interpreting human faces, as it has been mentioned in Chapter 6 already, these faces have to be depicted highly realistic to create the impression of a realistic scenery.

In this chapter, we focus on a selection of example applications that have been realized relying on the methods presented in previous chapters. These applications have been implemented partially in close cooperation with other disciplines and institutes. The first application is a quick example how face model fitting might be directly applicable in human-machine communication. The second and third application are linked, with the third building upon the second, and are inspired from ”facial mimicry”, an effect known from psychology.

7.1 Problem Statement

Research on automated facial expression recognition is mostly conducted without a specific application on mind and conducted on artificial image data. However, integrating facial expression recognition in running systems in real-world conditions raises different challenges. Obvious challenges are frame rate, lighting conditions, unpredicted human behavior or process communication. However, apart from such technical challenges, also the question how to evaluate such a system is raised. While technical measurements like accuracy and speed are still deducible by recording the interaction and labeling the recordings, inspection of the human factor requires different methods. To determine whether humans find the interaction with a system convenient or how humans react to a machine that recognizes facial expressions is difficult, at least, to determine from recordings.

7.2 Solution Idea

An established method to determine the opinion of humans on some fact are questionnaires. Psychologist are using this tool of evaluation since decades and therefore, we established cooperation with psychologists to utilize this method. We combine facial expression recognition with a facial expression synthesis component to create a system that responds to facial expressions in a way that is more intuitive than numbers or bar plots, especially for technically inexperienced person who are not familiar with automated facial expression recognition. This allows for creating an experimental setup that attracts a large variety of people, which is important to obtain unbiased surveys.

7.3 Related Work

Several authors have proposed the face as an alternative communication device in human-machine interaction already. Breazeal et al. present the robotic head ”Kismet” that depicts facial expressions depending on its ”emotions”, which are represented by several internal, motivational states [12]. This robot uses microphones to determine emotion in the user’s voice, but does not rely on the user’s facial expression. A system that not only displays but also recognizes facial expressions has been proposed by Bartlett et al. [6]. Their system recognizes seven basic emotions and depicts the recognized emotion on a virtual avatar. However, they do not include an emotion model, as Breazeal et al. did, and evaluation is given on the CK database only, not via a user study. Furthermore, facial expressions that are not directly related to the basic facial expressions, like yawning, are not considered in this approach. A similar approach is presented by Tscherepanow et al. [134]. Their approach is not bound to a fixed set of facial expressions, but they determine single motor commands directly from the visible image to mimic the facial expression. Unfortunately, they do not provide an evaluation of their system. Some researchers inspected facial expressions on robots or humans in dialogs, but these facial expressions are either depicted only by the robot or recognized in a wizard-of-oz setup [106, 6, 91]. The work presented in this thesis summarizes several of our earlier publications and research, conducted in two major stages. The first stage consisted of detecting the human facial expression to mirror it on a robot head and we refer to our earlier work for a more detailed description [16, 127]. The second stage integrated this in a human-machine dialog and included an evaluation in cooperation with psychologists [44].

Some applications are clearly industry driven, such as the smile detection of Whitehill et al. For their work, achieving high robustness is paramount since they aim to integrate it in digital cameras, where smiling should serve as a trigger to take pictures [55]. They aim at building a robust smile detection with the idea to integrate it in digital cameras, where smiling should serve as a trigger to take pictures. They create a very large database of training image to obtain this robustness. Shergill et al. propose to utilize facial expressions in marketing [124]. They observe that customers are discriminable in two categories: Customers that visit a physical or virtual store to see the supply without the intend to actually buy something and customers that enter the store with the proposition to do shopping. Shergill et al. argue that facial expressions are an applicable tool to distinguish these groups. Dhall et al. present a system that collects images based on the similarity of facial expressions in a gallery to simplify its browsing [27]. A similar idea is proposed by Kemelmacher et al., who determine the image from a gallery that matches a given test image best with respect to the person's head pose and facial expression [65].

Head gestures, such as head shaking and nodding, are simple and efficient communication instruments that are used by humans frequently in everyday life. Therefore, integrating them allows for a fast communication also with technical systems. Morency et al. present a system that integrates head gestures in traditional computer work like document browsing [89]. Another idea that is inspired by a medical application is the integration of head gesture recognition in intelligent wheel chairs for people who suffer from parkinson or quadriplegia [58]. We proposed a combined framework for head and hand gesture recognition in our earlier work [139]. The part of it concerned with head gesture recognition is detailed in this thesis, as well.

7.3.1 Conclusion

Two conclusions are drawn by many research groups: Integrating emotional feedback in human-machine interaction is a valuable goal and the face is a strong human communication tool. However, none of the mentioned systems provides combined facial expression recognition and synthesis that has been evaluated outside standard databases. Therefore, estimating the real benefit of facial expressions on the human-machine interface is still open research. System evaluations are usually focused on one aspect, either determining the technical parameters or the impact on the user. However, both aspects are important to obtain an impression of the system benefit.

7.4 Communication Framework

The applications presented in this chapter are composed of several components, each fulfilling a different task in the work chain. Designing the complete system as a collection of collaborative modules holds several advantages: Modules can be added and removed during runtime without having to reset the complete system, which is important, if one of the components fails for some reason. Furthermore, modules are simple to reuse in new projects. Examples of such modules are components that capture the camera data, processing units that extract features from these images, classification modules, for instance to extract head gestures from the features, and modules that control the robotic head. A fast and robust communication framework provides data transfer between these modules. We integrate the so-called ”Real-time Database”(RTDB) for this purpose, which has originally been developed for cognitive autonomous vehicles. As Goebel et al. demonstrate, the RTDB is capable of dealing with large amount of data input streams of different sources with different properties (i.e. data rate, packet losses, etc.) [43, 129]. Although, the RTDB is named ”database”, it is more a shared-memory implementation that provides ”write” and ”read” methods on a publisher-subscriber basis. This allows different software-modules the parallel access to the same input data without any blocking effects.

7.5 Head Gesture Recognition

Head gestures are a fast and convenient way to show agreement or disagreement in everyday communication. Head shaking and nodding are efficient communication signals, which we obtain in real-time also from camera images. The system extracts a 3D trajectory of the human head and then uses sequence-based classification to determine a head gesture from it. To determine the trajectory, a 3D face model is fit to the human's face and tracked through subsequent images. Since no facial expression information is required, we utilize a rigid face model to increase the frame rate. The pose information is transmitted via the RTDB to a sequence-classification module. Classification is performed via Hidden Markov Models, which have proven to be a quick and real-time capable classification method.

7.5.1 Feature Extraction

Since head gestures are inherently dynamic, we extract head movement rather than head position. The parameter vector of the rigid 3D model describes the model pose in space. For each image we calculate the head movement from the pose difference to the precedent image by $\boldsymbol{p}_t - \boldsymbol{p}_{t-1}$. Since we apply sequence classification rather than training classifiers for single images, we calculate a sequence of head movements with a sliding window approach over the last N images. The buffering capability of the RTDB has proven very helpful in doing so. The Hidden Markov Model is presented the feature vector $(\boldsymbol{p}_t - \boldsymbol{p}_{t-1}, \boldsymbol{p}_{t-1} - \boldsymbol{p}_{t-2}, \boldsymbol{p}_{t-2} - \boldsymbol{p}_{t-3}, \ldots, \boldsymbol{p}_{t-N} - \boldsymbol{p}_{t-N-1}$

7.5.2 Evaluation

We collect a set of image sequences depicting head shaking and nodding, consisting of 20 sample image sequences per class. Furthermore, a neutral head gesture reflects no specific head movement. We perform a stratified cross-validation with five folds to evaluate our approach. The obtained recognition results are presented in Table 7.1.

7.6 Facial Expression Mirroring on a Robot Head

The second application presented in this chapter displays the facial expression determined from camera images on a robot head. More specifically, this system

Classified as	Sequence Label		
	Shaking	Neutral	Nodding
Shaking	95%	5%	0%
Neutral	5%	85%	10%
Nodding	0%	0%	100%

Table 7.1: This table presents recognition rates of a HMM-based classification for the head gestures. The results are obtained from a five-fold cross validation.

determines the activation intensity of several FACS action units and mirrors them on the robotic head, allowing the robot to mirror the human's facial expression. Please note that the facial expression is not interpreted as being one of Ekman's universal facial expressions and therefore, the system is not limited to these facial expressions. The motivation for this experiment is based on insights in psychological research.

As already mentioned, facial expressions play an important role in human-human interaction. Humans that perceive emotional facial expressions on other humans' faces mirror these facial expressions within a few seconds, an effect known as ”facial mimicry”. Recent research investigates, whether this reaction is caused by the activation of so-called ”mirror neurons” that have been proven to be existent in the brain of monkeys and the human brain [107, 103]. There is evidence that observing an emotional facial expression does not only cause the observer to display the same facial expression but also induces the same emotion in the observer [36]. Furthermore, feeling empathy for others is also connected to the mirror neuron system [24, 40, 49]. Therefore, facial mimicry and empathy are linked, since facial mimicry signals a feeling of empathy and might induces it in the mirrored human, as well. This, in turn, can cause another mimicry reaction with switched roles, building a feeling of social bonding over time.

An open research question is, whether this facial mimicry effect is reproducible on the human-machine interface. The first experiment described in this chapter will provide the basis required to answer this question with the help of a robotic head. Obviously, if the facial mimicry effect should be evoked, the core assumption is that humans perceive the robot head's facial expression close to a corresponding human facial expression. For instance, if the human raises the eyebrows, than the robot's eyebrows should be perceived raised by the human as well, see Figure 7.1 for our experimental setup. Since it is difficult to evaluate this during runtime, we conduct an off-line evaluation of the system. The core

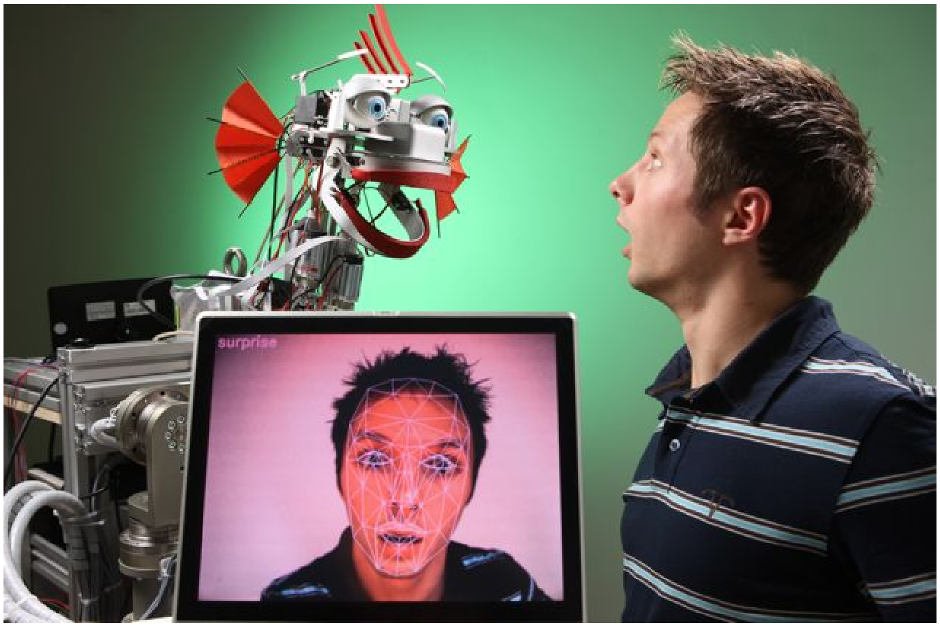


Figure 7.1: Demonstration setup with EDDIE (foto by Kurt Fuchs).

idea of this evaluation is the development of a similarity metric of the mirrored expression. This is achieved by conducting a user study to evaluate the degree, to which facial expressions displayed by the robot head and humans are perceived matching by untrained observers. The following experiment will then focus on the facial mimicry effect itself.

7.6.1 Model Tracking and Facial Action Units Analysis

Both, the model tracking and the FACS analysis, rely on a neutral reference image of the user. Since no previous information is available about the image content (except for the fact that a face is visible in it) or about the person in front of the robot head, model fitting is applied to determine reference model parameters $\boldsymbol{p}_0$. These reference parameters are specific for this person and are recalculated when the system is confronted with a new person. The model fitting is conducted in a two-step approach: Firstly, we fit a 2D shape model as described in Section 5.6 to determine the position of the face eyes and the face contour. From this information, the face pose in 3D space is calculated. Then, in the second step, the face shape parameters are calculated as described in Section 5.5.

In subsequent images, the face model is tracked to determine the model parameters $\boldsymbol{p}_t$. The face vertex points $\boldsymbol{v}$ are projected onto the image plane using perspective projection $\boldsymbol{j}$ to obtain their corresponding pixel coordinates $\hat{\boldsymbol{v}} = \boldsymbol{j}(\boldsymbol{v})$ in the camera image. Applying this to the reference frame $\mathcal{I}_0$ and the reference model parameters $\boldsymbol{p}_0$, we create pixel positions $\hat{\boldsymbol{v}}_0$. To estimate the corresponding pixel positions in the current frame $\mathcal{I}_t$ captured at time step t, we apply an optical flow method on $\mathcal{I}_0$ and $\mathcal{I}_t$ to calculate $\bar{\boldsymbol{v}}_t$. Afterwards, model parameters $\boldsymbol{p}_t$ are approximated that minimize the error between $\hat{\boldsymbol{v}}_t - \bar{\boldsymbol{v}}_t$. To calculate the FACS action unit activations, we rely on a subset of $\boldsymbol{a}$ that refers to the FACS system. We extract $\boldsymbol{p}_t^{FACS}$ from $\boldsymbol{p}_t - \boldsymbol{p}_0$ by selecting model parameters that refer to Action Units that are synthesizable by the robotic head. Therefore, $\boldsymbol{p}_t^{FACS}$ is a subset of the $\boldsymbol{p}^{FPI}$ features.

7.6.2 Facial Expression Synthesis

For the facial expression synthesis, the robot head EDDIE is used. EDDIE has been developed by Stefan Sosnowski at the ”Institute of Automatic Control Engineering” (LSR), at the TU München and has been generously provided by him for the conduction of this experiment. EDDIE is an emotion display with 23 degrees of freedom, mixing anthropomorphic and zoomorphic features [68]. With this head, 13 out of the 21 emFACS action units can be displayed .

The emotional state of the display can be controlled in two ways, either referring to the discrete basic emotions found by Ekman et al. or referring to the circumplex model of affect proposed by Russel et al.[31, 56]. Each state of the discrete basic emotions was modeled according to the emotion to FACS mapping by Ekman. With the robotic facial features being closely linked to the corresponding action units, a linear transformation from action units activation levels to the joint-space is used. This linear transformation is also used for the combined setup, were the activation of action units is directly provided by the analysis module. Transitions between states are animated by linear interpolation of the respective motor commands of the start - / end-state.

7.6.3 Experimental validation

The goal of this experiment is to evaluate if humans perceive the robotic facial expression close to a corresponding human facial expression. Since this is difficult during runtime, we extract static facial expression images from the CK database that has already been used in Chapter 6. The idea of the experiment is, to have the robot head mirror the facial expression visible in the example images, and have human raters compare the original human facial expression and the mirrored robot facial expression afterwards.

To determine the facial expression in these example images, we fit the face model to them and extract the feature vector $\boldsymbol{p}_t^{FACS}$. These feature values are then provided to the facial expression synthesis module to have the robotic head depict the facial expression. Since it is a known property of the database that the first image of each sequence depicts a neutral face, we rely on them for the calculation of $\boldsymbol{p}_0^{FACS}$ to support the calculation of $\boldsymbol{p}_t^{FACS}$. In total, 21 pictures are taken from the image database to determine the activation of the AUs with the facial analysis module. From this procedure we gain pairs of images, with one image depicting a human face and the corresponding second image depicting the robotic head mimicking the human face, see Figure 7.2 for an example. Since the data is automatically extracted and displayed by the system, as it would be in the mirror setup, this evaluation is a benchmark of the mirroring. This is done the same way as an actual video stream in the live mirroring setup is processed. The action units recognized by the analysis components and synthesized by the robot are AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU7 (lid tightener), AU13 (lip corner depressor), AU26 (yaw drop), AU42 (eyes closed).

7.6.4 Experiment Realization

To conduct the user study, a set of powerpoint slides with automatic data logging was created. Participants got the verbal instructions beforehand to follow the instructions on the screen and that they could work without a time limit. Twenty persons, six female and fourteen male, contributed to the evaluation. Since none of the persons is specifically trained on facial expression recognition or FACS coding, we decided against asking them to rate activations of specific action units. Instead, they were asked to rate human faces in four categories ($EyeBrows$, $EyeLids$, Jaw, $LipCorners$) and in five intensities. Example annotation were shown to the participants to prevent wrong labeling due to misunderstanding of the instructions, see Figure 7.3 (left) for the $EyeLids$ rating instruction. For the categories $EyeBrows$, $EyeLids$ and $LipCorners$, a low intensity represented lowered eye brows, closed eyes or depressed lip corners respectively. In conclusion, a high intensity reflected raised eye brow, wide opened eyes or raised lip corners. For the Jaw category, a low intensity referred to a closed mouth and a high intensity to a wide open mouth.

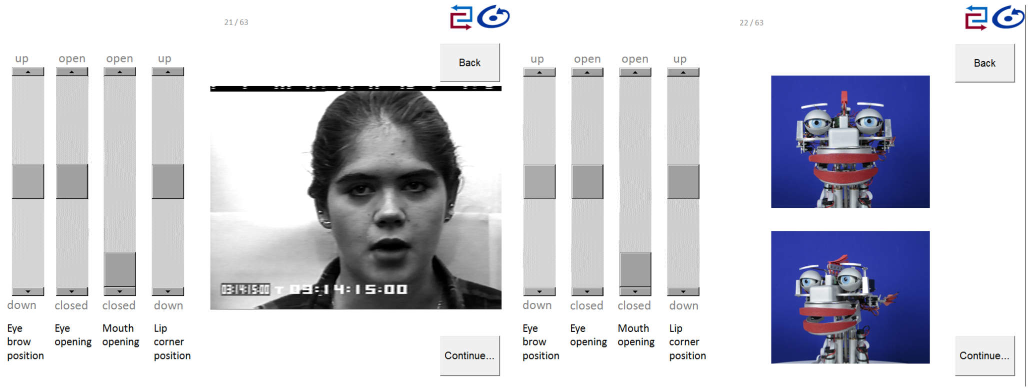


Figure 7.2: We create a set of image pairs with one image depicting a human and a second image depicting the robot head mimicking the human. The participants were not aware that they were presented pairs of images.

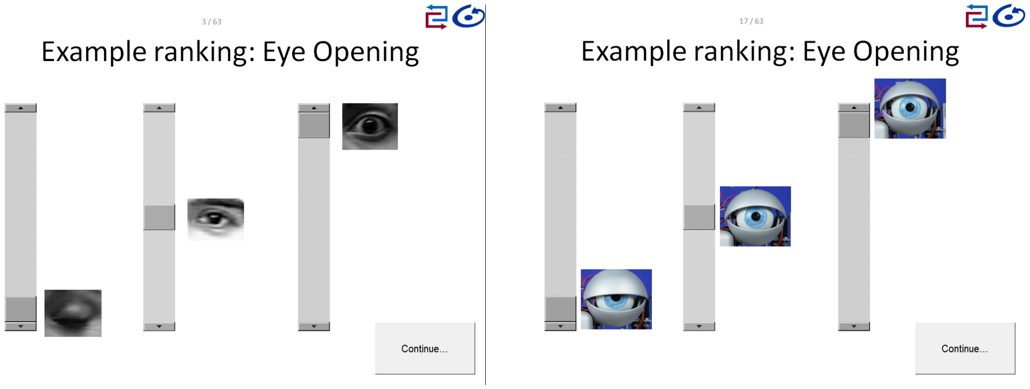


Figure 7.3: Participants were asked to rate human and robotic faces in four categories and five intensities.

In the first evaluation phase, the participants were presented eight images depicting human faces in a predefined order. These images did not have corresponding images with the robotic head. The first phase served two reasons: First, to ensure that the participants had correctly understood the task, second, to have a reference of the users' ability to rate facial expressions. In the second evaluation phase, 21 images from the image database and 21 corresponding images of the robotic head were presented to the participants in randomized order. Please note that the participants were not informed that the image data includes matching human-robotic head pairs. The order was different for every participant. Figure 7.2 depicts an example of a human-robotic head pair without rating, i.e. all sliders are in initial state. Similar to the first phase, participants were informed on an introduction slide that a robotic head would now be depicted as well and example ratings were given, see Figure 7.3 (right). There was no difference in the rating mechanisms for human faces and robotic heads, except for the fact that two images were presented for the robotic head. Participants were allowed to navigate freely through the test with continue/back buttons.

7.6.5 Results

In this section, we inspect the finding of the experiment conduction. First, we inspect the similarity of participants' rating schemes. If different participants applied a different rating scheme, their rating would not be comparable, although they might actually have the same perception of the activation of certain action units. As mentioned in Section 7.6.3, the participants were presented eight training image with four sliders each, resulting in 32 slider values. The mean variance for all slider values is 0.41, which demonstrates that participants rated the training images very similarly and therefore their rating is comparable.

A similar idea is applied to inspect the consistency between the rating of a human face and a corresponding robotic face. We denote the rating of the human face in one of our $L = 21$ image pairs by one of the $N = 20$ participants with $h_c^{l,n}$ with $1 \leq l \leq L, 1 \leq n \leq N$ and $c \in \{EyeBrows, EyeLids, Jaw, LipCorners\}$. The ratings of the robotic face are denoted by $r_c^{l,n}$. Per participant $21x4x2 = 168$ (21 image pairs, 4 sliders, 2 images per pair) values for all image pairs l, and all categories c are calculated. To inspect the similarity between human and robotic face rating, we calculate $e_c^{l,k} = h_c^{l,n} - r_c^{l,n}$. Furthermore, to group our inspection by category, we create data vectors $\boldsymbol{e}_X$ that contain all values $e_{c=X}^{l,k}$.

We calculate a histogram of $\boldsymbol{e}_{EyeBrows}$,$\boldsymbol{e}_{EyeLids}$,$\boldsymbol{e}_{Jaw}$ and $\boldsymbol{e}_{EyeBrows}$ to obtain an intuition of the rating discrepancy distribution, see Figure 7.4. For all categories the most frequent value of $\boldsymbol{e}_c$ is 0, which indicates that participants rated the robotic head and the human face equally. Furthermore, only a very small fraction of values of $e_c^{l,k}$ has $e_c^{l,k} < -1$ or $e_c^{l,k} > 1$, which leads to the conclusion that ratings of the human face and the robotic head only rarely differ more than one slider unit.

To obtain comparable numbers, we calculate mean and variance of $\boldsymbol{e}_{EyeBrows}$, $\boldsymbol{e}_{EyeLids}$, $\boldsymbol{e}_{Jaw}$ and $\boldsymbol{e}_{EyeBrows}$, see Table 7.2 and Figure 7.5 for a visualization. Inspecting the small mean values confirms our findings. Larger values would imply a general shift between the human and robotic head. For instance, a high

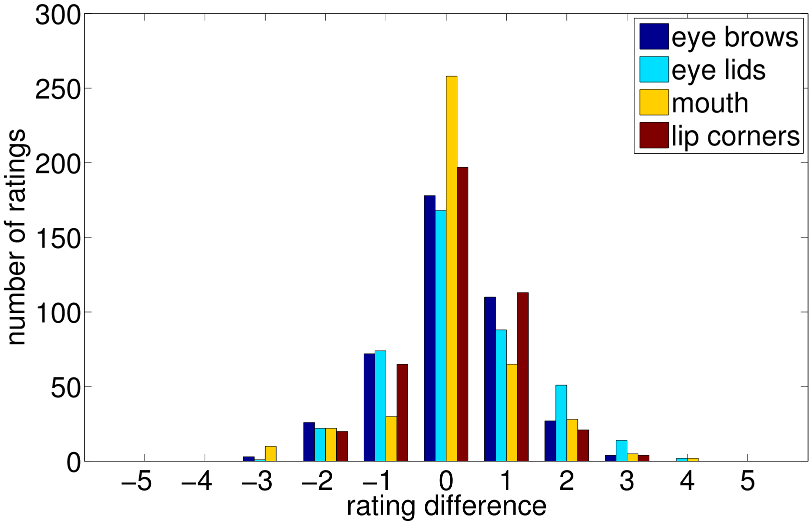


Figure 7.4: Only a small fraction of rating differences between a human face and a corresponding robotic head are larger than one slider unit.

mean value at Jaw would indicate that the robotic head's yaw drop is always perceived smaller than the human one. Furthermore, the variances for $EyeBrows$, Jaw and $LipCorners$ are all close to 1.0, which further strengthens our findings that ratings of the human face and the robotic head only rarely differ more than one slider unit in these categories. $EyeLids$, however, shows a larger variance, which indicates that this has been the most difficult category to rate by the participants. However, since the variance is still less than 2.0, the rough status (eyes opened/eyes closed) still has been recognized correctly in general.

category	mean	variance
Eye Brows	0.10	1.08
Eye Lids	0.28	1.40
Jaw	0.10	1.10
Lip Corners	0.15	0.88
overall	**0.16**	**1.11**

Table 7.2: Mean and variance in the rating difference of a human face and corresponding robotic head.

7.7 Facial Expressions in a Human-Robot Dialog

Inspired by the insights that humans perceive the facial expression mirrored on the robot close to the original, human facial expression, we integrate the system in a running dialog. The idea of this experiment is two-fold: Firstly, to reproduce the empathy inducing effect of facial mimicry and secondly to demonstrate that human-robot interaction benefits from the integration of facial expression analysis and synthesis in the interaction process. Users are asked to play a game of ”Akinator” with the robot head, while the robot head is in one of three states: The robot head either ignores the human’s facial expression completely, or simply mirrors the human’s facial expression, or determines its own reaction from a social model that takes the robot’s internal state and the user’s facial expression into consideration. We now shortly introduce the components that have been added in comparison with the last experiment: The ”Akinator” and the social model.

7.7.1 Dialog and Akinator

To create a backbone for the ongoing dialog, an interface to the ”Akinator” (see www.akinator.com) is integrated. This web-based application, which is usually executed in a browser, realizes a simple game, in which the ”Akinator” tries to guess a person chosen by the user. In our application, the experiment participant takes the role of a user and is asked to choose a person. The person may be a real or

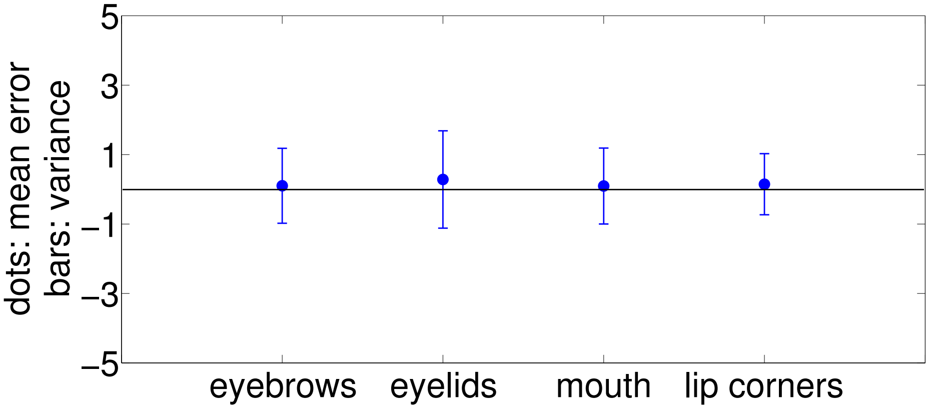


Figure 7.5: Mean errors in rating of human versus mirrored robotic expressions grouped by action units

fictional person, currently living or historical, taken from literature, the media or public live. Then, EDDIE tries to guess the person by asking several questions and using its interface to the ”Akinator” to transmit the user’s answer and to obtain the next question. To answer Akinator’s questions, a set of fixed answers is presented by the system. The set of answers is the same for every question and consists of: ”Yes”, ”Probably” / “Partially”, ”I don’t know”, ”Probably not”/ ”Not really”, and ”No”. Example questions asked by the Akinator are: ”Is your character a girl?”, ”Does your character live in America” or ”Does your character really exist?”. Since it is the idea of this experiment to create the illusion that the participant is in a dialog with EDDIE, traditional interface methods like mouse and keyboard are not suitable. Instead, text-to-speech is used to present Akinator’s questions acoustically to the participant and speech recognition is utilized to determine the participant’s answers.

The interfaces and their interplay, including the text-to-speech subcomponent, the speech recognizer and the communication to the ”Akinator” interface have been provided by Jürgen Blume from the Institute for Human-Machine Interaction of the Technische Universität München. We would like to acknowledge his contribution to the realization of this experiment.

7.7.2 Social Model

The utilized model is based on a reduced version of the ”Zurich Model of Social Motivation” [46], which has again been provided by Stefan Sosnowski from the Institute of Automatic Control Engineering of the Technische Universität München. We provide only a short overview of the model here, but refer to Borutta et al. for a more detailed description [11].

The model has been developed to describe the behavior of children in the presence of other humans. In our implementation, both, the child and the human are situated in a virtual environment. The child’s emotional state is modeled by three interrelated subsystems: The autonomy system, the arousal system and the security system. The child will try to achieve a situation, in which these three subsystems are balanced. Actions occur due to changes in any of these subsystems that disturb this balancing. Facial expressions are considered reactions to changes in one of the subsystems and the subsequent adaption process [11]. For instance, smile reactions are the result of a decline in autonomy like social distance changes, environmental changes or conflicts.

Humans in the environment are described with three parameters: The familiarity, the relevancy and the distance to the child. The familiarity models the degree, to which the child is familiar with the human. A high value would for instance be assigned to the child's mother, whereas a small value would be a complete stranger. The relevancy models the strength of the impact the person's presence has on the child. A small value would indicate for instance a peasant who obviously ignores the child. An example for a high value is the child's doctor. The distance is calculated from the child's position and the human's position in the virtual environment. Humans produce a potential of security and arousal in their surrounding, depending on their settings of familiarity and relevancy. Depending on its internal state, the child moves in the environment to find a place that supports its current need for arousal or security.

In our application, the robot takes the role of the child and the participant takes the role of a human with high relevancy and medium familiarity. Since the physical robot is fixed on the table, we immobilize the virtual child, as well, so that it can't move either. The human's position in front of the camera is obtained from the face model pose and mapped to a position in the virtual environment. If the participant moves, this influences the robot's emotional subsystems, due to the generated potential of arousal and security. For instance, it is possible to "scare" the robot by quickly moving towards it. This produces a large amount of arousal, which results in a surprise reaction of the robot. Further changes in the subsystem are induced by detecting the human's facial expression in this experiment. Thereto, we extend the model with an interface to our facial expression recognition system. In contrast to the previous experiment described in section 7.6, not only the activation of single action units is detected, but a classifier is trained as described in Chapter 6 to detect the intensity of certain facial expression. We utilize support vector regression that is trained on the $\boldsymbol{p}^{FPI}$ features to determine the intensity of smiling or surprise from the user's face. The required neutral image is easily obtained in the beginning of the dialog. We model smiling at the robot to increases its security state, which corresponds to the idea of providing security to the robot. If the robot is in a balanced state, this results in a smiling back reaction of the robot. Detected surprise increases the arousal level of the robot, which corresponds to signaling the robot that something exciting has just happened, thus also "exciting" the robot.

7.7.3 Experiment Conduction

Experiment participants were seated in a quiet room with controlled lighting for the experiment. They were grouped in three different groups referring to the condition of the robot head, which was either not reacting to the participant's facial expression, mirroring the facial expression or depicting its facial expression induced by the social model. Participants were not informed about the goal of the experiment beforehand and were only told that the experiment was about human-robot interaction. Then they were instructed on the ”Akinator” game and asked to pick a person of their choice for the game. Immediately after the game, they were asked to fill in a randomized questionnaire, which is presented in Table 9.3 in the Appendix. The goal of the experiment was to inspect whether the chosen condition influenced the user’s perceived empathy towards the robot, the performance of the robot as it is subjectively perceived by the user and the user acceptance of the robot. Please note that we did not measure the objective performance of the robot here, which would be for instance the time it took the robot to guess the person or the number of questions. The questionnaire and its evaluation has been provided by Barbara Gonsior from the Institute of Automatic Control Engineering of the Technische Universität München.

The questionnaire has been designed to measure user acceptance in five different categories:

- *Trust*: The belief that the system performs with personal integrity and reliability.

- *Perceived Sociability*: The perceived ability of the system to perform sociable behavior.

- *Social Presence*: The experience of sensing a social entity when interacting with the system.

- *Perceived Enjoyment*: Feelings of joy or pleasure associated by the user with the use of the system.

- *Intention to Use*: The outspoken intention to use the system over a longer period in time.

To determine the user’s empathy and perceived performance, the questionnaire has been split in two parts, depending on whether EDDIE was successful in guessing the person or not. Participants answered the questions on a Likert scale with five intensities from 1 (strong disagree) to 5 (strong agree). Some of the questions were negated versions of other questions. The rating of these questions had to be negated, too, during evaluation.

category	**Condition**		
	Neutral	Mirrored	Social Model
Empathy	3.1(1.3)	3.7(1.1)	4.4(0.8)
Subjective Performance	2.8(1.2)	3.4(1.0)	4.1(0.9)
Trust	3.0(0.6)	3.3(0.8)	3.7(0.5)
Perceived Sociability	3.2(1.0)	3.6(1.0)	3.9(0.7)
Social Presence	2.8(0.6)	2.8(0.7)	2.9(0.7)
Perceived Enjoyment	2.8(1.4)	3.9(1.2)	4.2(0.7)
Intention to Use	3.0(1.3)	3.5(1.0)	3.9(1.0)

Table 7.3: All categories show a clear tendency that ratings increase with the condition. This tendency is strongest with ”subjective performance”, ”empathy” and ”perceived enjoyment”.

7.7.4 Results

As Table 7.3 shows, the categories involving facial expressions are rated higher, i.e. ”better” from the user’s point of view, than the neutral condition. This is specifically true for the ”empathy” condition, which is important, because this indicates that the behavior of the robot indeed evokes a feeling of empathy or sympathy for the robot. Therefore, participants reacted on the robot’s facial mimicry. This also correlates with the category ”perceived enjoyment”, which indicates that participants also enjoyed playing with EDDIE more, when it reacted to their facial expressions. The category ”subjective performance” together with the ”intention of use” category shows, that integrating facial expressions in the process improves the user’s impression of the machine performance and provides strong motivation to use or reuse the machine. People feel more satisfied with the machine performance afterwards, independent of the actual or objective performance of the machine, than with an emotion-neutral interface, and are more likely to come back and use it again. The difference between the conditions ”Mirrored” and ”Social Model” further indicates, that modeling the machine behavior itself still has impact on the impression the machine provokes in the user. It is not only beneficial to integrate facial expression recognition, but also to model the machine agent and its internal state itself.

7.8 Discussion

In this chapter, we depicted example applications that have been realized with the techniques presented in earlier chapters. We shortly introduced head gestures for intuitive human-machine communication and integrated facial expression recognition with a robot head.

Although the experiment with the robot head might look like a nice toy for engineers and computer scientists at first glance, it provided the opportunity to gain interesting research insights. Although much research is conducted on facial expression recognition for human-robot interaction, its beneficial effect is usually axiomatically assumed. Only little work is dedicated to inspecting the actual benefit. Our evaluation proofs, that facial expression recognition is beneficial to human-robot-interaction, an insight, from which a large community of researchers will benefit, as well. However, this insight also raises further research questions that still need to be answered, for instance, whether other reactions than empathy are inducible, as well. Scientifically speaking, the idea to create a robot that is able to induce anger in humans would be interesting, as well, as it would demonstrate that the robot is perceived more as a person than as a machine. This again raises the question, what is actually required for humans to perceive machines as a social person rather than a lifeless object. Facial expressions seem to be part of the answer.

Chapter 8

Discussion and Future Work

Whitehill et al. recently recognized one of the major challenges that research on facial expression recognitions systems is currently facing. Traditionally, facial expression recognition systems are evaluated on standard databases. They offer images of facial expressions that are acted according to the instructions of the database authors. Whitehill et al. refer to these databases when they state that ”It is conceivable that by evaluating performance on these data sets the field of automatic expression recognition could be driving itself into algorithmic ’local maxima’. ” [55]. In this thesis, we identified one of the driving factors of the dangerous tendency to be the evaluation strategy that is typically applied to test facial expression recognition algorithms. Mostly self-classification evaluation, like percentage-split, stratified cross-validation or leave-one-out validation is applied. The reason for using the aforementioned databases is that obtaining real, non-acted data is difficult and ethically doubtful, especially for facial expressions displaying emotions like fear or sadness. However, this approach induces bias in the database data due to the author’s instructions to subjects. This effect is aggravated by the fact that many approaches consider apex facial expressions (facial expressions with maximum intensity) only. This danger becomes evident when conducting cross-database evaluation instead of self-classification evaluation. It has been found by several research groups, that results in cross-database evaluation are significantly below self-classification evaluation, these findings have not sparked a major change in evaluation strategies so far. However, we concluded that we can not expect algorithms to work robustly in real-world applications if they do not even generalize to data taken from another database. The benefits of this evaluation strategy are that it prevents classifiers from specializing on database properties and allows for a more realistic comparison of classifiers, and thus wards against getting caught in a local maximum. Therefore, we inspected the benefit of cross-database evaluation in more detail in order to demonstrate its contribution to conducting meaningful evaluations.

The second danger in the tendency mentioned by Whitehill et al. is that it results in a gap between the facial expressions that recognition algorithms are trained on and facial expressions as they appear in the real world. This is unfortunate, because practical application of facial expression recognition offers many interesting research opportunities in human-machine interaction. There is no doubt, that facial expressions are important elements of human communication. However, although its benefit in human-machine communication is usually assumed, the question whether they have a similar, or at least significant, impact on human-machine communication is still open. We approached this question and proposed an answer by conducting an experiment in which a human participates in a dialog with a robot head that itself depicts facial expressions and reacts to the human's facial expression. A survey conducted in close cooperation with psychologists revealed that participants perceive the robot to work more effective and enjoy the cooperation more when it reacts to their facial expression. The answer to this question is of interest to the community, since it provides a strong motivation for research in this area, not only as a theoretical pattern recognition problem but with practical application in mind.

Our application required a face model that represents semantic information about facial actions, like rising the eyebrows, in single parameters. Unfortunately, face models like ASMs and AAMs are generated from statistics in manually labeled training data and their model parameters refer to statistical variances in the annotation rather than semantic face movements. Fortunately, the Candide-III face model provides model parameters with semantic interpretation. Since model fitting strategies are usually tied to and only applicable with a specific type of model, we proposed a novel fitting strategy, as well. Directly comparing two representatives of major model fitting categories that are widely used, those utilizing objective functions and those utilizing displacement expert, the later turned out to be faster and more robust. Since the Candide-III model parameters refer the shape and movement of specific facial components, we integrate an image representation that specifically highlights these components. However, this image representation is not bound to the Candide-III face model, but is embeddable into other fitting strategies, as well.

We stated three contributions of this thesis in the introduction. We will now visit them again to verify how they have been fulfilled.

We presented a system for facial expression recognition from camera images.

The proposed approach processes images in three subsequent steps: A preprocessing step that generates multi-band images from the raw image data, a model fitting step that is applied to the publicly available Candide-III face model, and finally a classification step that determines the facial expression from the model parameters. Each of these steps works automatically and no manual interference is required.

Furthermore, we presented evaluations for implementations of those steps that are tuned to reflect robustness in real-world scenarios by integrating separate training and test databases. Our preprocessing method is specifically tuned to segment the face from the image background and the facial components from the rest of the face. We demonstrated that model fitting greatly benefits from integrating this novel image representation in the fitting process, which is analyzed on the "Labeled Faces in the Wild" database. This database depicts images collected from the media that offer a large variety with respect to age, ethnic background, clothing style, background and head pose. The classifiers that determine the facial expression have been evaluated following a novel evaluation strategy in cross-database scenarios to ward against overspecialization.

Finally, we presented applications of our algorithms to demonstrate several possible applications of the presented techniques. We shortly introduced an application to head gesture recognition, which is a convenient and efficient human communication modality. Then, we integrated facial expression recognition in a human-robot dialog. Apart from being a platform to demonstrate facial expression analysis and synthesis, it provided a research platform for emotions in human-machine interaction.

Future work will consider integrating multi-band images not only in face model fitting but also in face model tracking. Two important step have to be taken to achieve this: Firstly, the computation of the multi-band images has be be conducted faster than real-time. Since the calculation has large potential for parallelization, the computation will be shifted to GPUs of modern graphics cards. Secondly, the process will be reformulated to take prior knowledge about the person visible in the camera images into consideration. This will be achieved by adding person-characteristics, similar to the image-characteristics, which will further speedup the process and increase its accuracy. Furthermore, we will integrate real facial expression data in our classifier training and evaluation. This includes two steps: assembly of a database from sources like media and evaluation of this additional database with our cross-database strategy.

Another large point of interest is the integration of more recent sensor hardware like the Kinect sensor. Although 3D data, for instance from laser scans, has been considered for facial expression recognition, practical application of such techniques have always been difficult due to the sensor restrictions. With the Kinect sensor, there is the opportunity to consider depth information in combination with camera images, synchronized, registered and in in real-time. It is to be expected that this information will be particularly helpful in fitting the 3D model. Preliminary experiment show that the head pose can be extracted in real-time from this data and that the initial pose estimation of the 3D model also benefits from it. This, in turn, will increase the fitting accuracy.

Inspecting the development in this area over the last decades reveals that face image analysis receives a steadily growing interest from the scientific community and industry alike. Facial expression recognition offers opportunities for advanced pattern recognition as well as fascinating interdisciplinary research. Applications already range from smile-shutters for digital cameras to face analysis for movie production, documenting the progress that has been achieved through research on the subject. However, some of the research questions that are still open were outlined in this thesis, providing room for continued and interesting work. We look forward to see the insights that researches all over the world will gain in the years to come.

Chapter 9

Appendix

AU	Name	Facial muscles involved
1	Inner Brow Raiser	*Frontalis (pars medialis)*
2	Outer Brow Raiser	*Frontalis (pars lateralis)*
4	Brow Lowerer	*Corrugator supercilii*, *Depressor supercilii*
5	Upper Lid Raiser	*Levator palpebrae superioris*
6	Cheek Raiser	*Orbicularis oculi (pars orbitalis)*
7	Lid Tightener	*Orbicularis oculi (pars palpebralis)*
9	Nose Wrinkler	*Levator labii superioris alaeque nasi*
10	Upper Lip Raiser	*Levator labii superioris*
11	Nasolabial Deepener	*Zygomaticus minor*
12	Lip Corner Puller	*Zygomaticus major*
13	Cheek Puffer	*Levator anguli oris*
14	Dimpler	*Buccinator*
15	Lip Corner Depressor	*Depressor anguli oris*
16	Lower Lip Depressor	*Depressor labii inferioris*
17	Chin Raiser	*Mentalis*
18	Lip Puckerer	*Incisivii labii superioris* and *Incisivii labii inferioris*
20	Lip stretcher	*Risorius* and *platysma*
21	Neck Tightener	
22	Lip Funneler	*Orbicularis oris*
23	Lip Tightener	*Orbicularis oris*
24	Lip Pressor	*Orbicularis oris*
25	Lips part	*Depressor labii inferioris* or relaxation of *Mentalis*, or *Orbicularis oris*
26	Jaw Drop	*Masseter*, relaxed *Temporalis* and internal *pterygoid*
27	Mouth Stretch	*Pterygoids* and *Digastric*
28	Lip Suck	*Orbicularis oris*
31	Jaw Clencher	
38	Nostril Dilator	
39	Nostril Compressor	
43	Eyes Closed	Relaxation of *Levator palpebrae superioris*, *Orbicularis oculi (pars palpebralis)*
45	Blink	Relaxation of *Levator palpebrae superioris*, *Orbicularis oculi (pars palpebralis)*
46	Wink	Relaxation of *Levator palpebrae superioris*, *Orbicularis oculi (pars palpebralis)*

Table 9.1: List of Action Units and their corresponding facial muscles, based on [35].

Facial expression parameters	
6	Upper lip raiser (AU10)
7	Jaw drop (AU26/27)
8	Lip stretcher (AU20)
9	Brow lowerer (AU4)
10	Lip corner depressor (AU13/15)
11	Outer brow raiser (AU2)
12	Eyes closed (AU42/43/44/45)
13	Lid tightener (AU7)
14	Nose wrinkler (AU9)
15	Lip presser (AU23/24)
16	Upper lid raiser (AU5)
Shape parameters	
17	Head height
18	Eyebrows vertical position
19	Eyes vertical position
20	Eyes, width
21	Eyes, height
22	Eye separation distance
23	Cheeks z
24	Nose z-extension
25	Nose vertical position
26	Nose, pointing up
27	Mouth vertical position
28	Mouth width
29	Eyes vertical difference

Table 9.2: The Candide-III face model parameters that model the face shape and the FACS action units.

	Empathy
1	I am happy that Eddie guessed my person. It's a shame Eddie didn't guess my person.
2	I would have been proud if Eddie hadn't guessed my person. I'm proud Eddie didn't guess my person.
3	It would have been a pity if Eddie didn't guess my person it would have been nice if Eddie had guessed my person.
4	It took Eddie long to guess my person. It took Eddie too long to guess my person.

	Subjective Performance
1	I was impressed by how fast Eddie has guessed my person.
	I had the feeling that Eddie nearly guessed my person.
2	Eddie has shown a good performance.
3	I think that Eddie has worked efficiently.
4	It took Eddie long to guess my person.
	It took Eddie too long to guess my person.

	Trust
1	I would believe Eddie if he gave me advice.
2	Eddie is inspiring confidence.
3	I feel that I can trust Eddie.
4	I do not trust Eddie's statements.

	Perceived Sociability
1	I like Eddie.
2	Eddies mimic and verbal statements fit together well.
3	Eddie was good conversation partner.
4	Eddie's behavior was inappropriate.

	Social Presence
1	I had the feeling that Eddie really looked at me.
2	I could imagine Eddie as a living being.
3	Sometimes it felt like Eddie had real feelings.
4	Eddies behavior was not humanlike.

	Perceived Enjoyment
1	It was fun to interact with Eddie.
2	The conversation with Eddie was fascinating.
3	I consider Eddie to be entertaining.
4	It's boring when Eddie interacts with me.

	Intention to Use
1	I would like to interact with Eddie more often.
2	I would take Eddie home with me.
3	I would like to play again with Eddie within the next few days.
4	I could imagine interacting with Eddie over an extended period of time.

Table 9.3: The questionnair used in the human-robot interaction experiment.

Bibliography

[1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1), jan 2006.

[2] J. Ahlberg. Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden, 2001.

[3] Jeffrey F. Cohn Tsuhan Chen Zara Ambadar Kenneth M. Prkachin Ahmed Bilal Ashraf, Simon Lucey and Patricia E. Solomon. The painful face pain expression recognition using active appearance models. *Image and Vision Computing*, 27, 2009.

[4] Keith Anderson and Peter W. McOwan. A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man and Cybernetics*, 36(1), 2006.

[5] Ioana Bacivarov and Peter M. Corcoran. Facial expression modeling using component aam models - gaming applications. 2009.

[6] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Proceedings of the Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, helt in conjunction with CVPR*, 2003.

[7] May Beerenbaum. Face time. *American Entomologist*, 51(2):68–69, 2005.

[8] Volker Blanz. Face recognition based on a 3D morphable model. In *Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG 2006), 10-12 April 2006, Southampton, UK*, pages 617–624. IEEE Computer Society, 2006.

[9] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In Alyn Rockwood, editor, *Siggraph 1999, Computer Graphics Proceedings*, pages 187–194, Los Angeles, 1999. Addison Wesley Longman.

[10] George Borshukov and J. P. Lewis. Realistic human face rendering for ”the matrix reloaded”. In *Proceedings of the SIGGRAPH*, 2003.

[11] I. Borutta, S. Sosnowski, K. Khnlenz, M. Zehetleitner, and N. Bischof. Generating artificial smile variations based on a psychological system-theoretic approach. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man)*, Toyama, Japan, 2009.

[12] Cynthia Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 2003.

[13] UJose M. Buenaposada, Enrique Munoz, and Luis Baumela. Recognising facial expressions in video sequences. *Pattern Analysis and Application*, 11(1), 2008.

[14] F.W. Campbell. How much of the information falling on the retina reaches the visual cortex and how much is stored in the visual memory? *Pattern Recognition Mechanisms*, pages 83–95, 1983.

[15] Yisong Chen and Franck Davoine. Simultaneous tracking of rigid head motion and non-rigid facial animation by analyzing local features statistically. In *British Machine Vision Conference*, 2006.

[16] Kolja Khnlenz Bernd Radig Christoph Mayer, Stefan Sosnowski. Towards robotic facial mimicry: system development and evaluation. In *International Symposium in Roboth-Human Interactive Communication*, 2011.

[17] Matthias Wimmer Christoph Mayer and Bernd Radig. Adjusted pixel features for facial component classification. *Image and Vision Computing Journal*, 2009.

[18] Tim F. Cootes, G. J. Edwards, and Chris J. Taylor. Active appearance models. In H. Burkhardt and Bernd Neumann, editors, *5th European Conference on Computer Vision*, volume 2, pages 484–498, Freiburg, Germany, 1998. Springer-Verlag.

[19] Tim F. Cootes and Chris J. Taylor. Active shape models – smart snakes. In *Proceedings of the 3rd British Machine Vision Conference*, pages 266 – 275. Springer Verlag, 1992.

[20] Tim F. Cootes and Chris J. Taylor. On representing edge structure for model matching. *Computer Vision and Pattern Recognition*, 1:1114–1119, March 2001.

[21] D. Cristinacce and T.F. Cootes. Facial feature detection and tracking with automatic template selection. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006.

[22] David Cristinacce and Tim F. Cootes. Feature detection and tracking with constrained local models. In *17th British Machine Vision Conference*, pages 929–938, Edinburgh, UK, 2006.

[23] David Cristinacce and Tim F. Cootes. Boosted regression active shape models. In *Proceedings of the British Machine Vision Conference*, volume 2, pages 880–889, 2007.

[24] M. Dapretto, M. S. Davies, J. H. Pfeifer, A. A. Scott, M. Sigman, S. Y. Bookheimer, and M. Iacoboni. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience*, 2005.

[25] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, London, United Kingdom, 1872.

[26] Fadi Dornaika Franck Davoine. Simultaneous facial action tracking and expression recognition in the presence of head motion. que, 2008.

[27] Abhinav Dhall, Akshay Asthana, and Roland Goecke. Facial expression based automatic album creation. In *International Conference on Neural Information Processing*, 2010.

[28] Liya Ding and Aleix M. Martinez. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32(11), 2010.

[29] Fadi Dornaika and Bogdan Raducanu. Three-dimensional face pose detection and tracking using monocular videos: Tool and application. 39, 2009.

[30] Guillaume-Benjamin-Armand Duchenne. *Mcanisme de la Physionomie Humaine o, Analyse lectro-physiologique de l'expression des passions*. J.-B. Baillire, Paris, France, 1876.

[31] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In J. Cole, editor, *Nebraska Symposium on Motivation 1971*, volume 19, pages 207–283, Lincoln, NE, 1972. University of Nebraska Press.

[32] Paul Ekman. Facial expressions. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*, New York, 1999. John Wiley & Sons Ltd.

[33] Paul Ekman, R. Davidson, and Wallace Friesen. The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology*, 58(2):342–353, 1990.

[34] Paul Ekman and Wallace Friesen. *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*. Consulting Psychologists Press, San Francisco, 1978.

[35] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager, editors. *Facial Action Coding System*. A Human Face, 666 Malibu Drive, Salt Lake City UT 84107, 2002.

[36] Laura J. Mann Eric J. Moody, Daniel N. McIntosh and Kimberly R. Weisser. More than mere mimicry? the influence of emotion on rapid facial reactions to faces. *Emotion*, 7(2), 2007.

[37] N. Eveno, A. Caplier, and P.Y. Coulon. Jumping snakes and parametric model for lip segmentation. In *Proceedings of the International Conference on Image Processing*, 2003.

[38] B. Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36, 2003.

[39] Wallace V. Friesen and Paul Ekman. *Emotional Facial Action Coding System*. Unpublished manuscript, University of California at San Francisco, 1983.

[40] V. Gallese. The 'shared manifold' hypothesis. *Journal of Consciousness Studies*, 8, 2001.

[41] Feng Ge, Song Wang, and Tiecheng Liu. Image-segmentation evaluation from the perspective of salient object extraction. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition*, 2007.

[42] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, June 1975.

[43] M. Goebl and G. Frber. A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles. *Intelligent Vehicles Symposium*, pages 737 – 740, June 2007.

[44] Barbara Gonsior, Stefan Sosnowski, Christoph Mayer, Jürgen Blume, Bernd Radig, Dietmar Wollherr, and Kolja Kühnlenz. Improving aspects of empathy subjective performance for hri through mirroring emotions. In *International Symposium in Roboth-Human Interactive Communication*, 2011.

[45] R. Gross, I. Matthews, S. Baker, and T. Kanade. The cmu multiple pose, illumination and expression database. Technical report, Robotics Institute Carnegie Mellon University, 2007.

[46] Harry Gubler and Norbert Bischof. A systems' perspective on infant development. in (eds.),(1-37). hillsdale: Lawrence erlbaum. *Infant Development: Perspectives from German-speaking Countries*, pages 1–37, 1990.

[47] Onur C. Hamsici and Aleix M. Martinez. Active appearance models with rotation invariant kernels. In *Proceedings of the International Conference on Computer Vision*, 2009.

[48] Trevor Hastie. Classification by pairwise coupling. *Advances in Neural Information Processing Systems*, 1998.

[49] Ursula Hess and Sylvie Blairy. Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology*, 40(2):129 – 141, 2001.

[50] Erik Hjelmasa and Boon Kee Lowb. Face detection: A survey. 83, 2001.

[51] C.-H. Hjortsj. *Mnniskans ansikte och det mimiska sprket*. Studentlitertur, Lund, Sweden, 1969.

[52] Kenny Hong, Stephan K. Chalup, and Robert A.R. King. A component based approach improves classification of discrete facial expressions over a holistic approach. 2010.

[53] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K. Jain. Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, May 2002.

[54] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments, 2008.

[55] Ian Fasel Marian Bartlett Jacob Whitehill, Gwen Littlewort and Javier Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2009.

[56] J.A.Russell. A circumplex model of effect. *Journal of Personality and Social Psychology*, 1980.

[57] Oliver Jesorsky, Klaus J. Kirchberg, and Robert Frischholz. Robust face detection using the hausdorff distance. In *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 90–95, Halmstad, Sweden, 2001. Springer-Verlag.

[58] Pei Jia, Huosheng H. Hu, Tao Lu, and Kui Yuan. Head gesture recognition for hands-free control of an intelligent wheelchair. *Industrial Robot: An International Journal*, 2007.

[59] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *Interational Journal of Computer Vision*, 46(1):81–96, 2002.

[60] Vlad Popovicib Julien Meyneta and Jean-Philippe Thiran. Face detection with boosted gaussian features. 40, 2007.

[61] Frédéric Jurie and Michel Dhome. Hyperplane approximation for template matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7), 2002.

[62] Fathi Kahmaran and Muhittin Gokmen. Illumination invariant face alignment using multi-band active appearance models. In *Pattern Recognition and Machine Intelligence*, pages 118–127, 2005.

[63] Takeo Kanade, John F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, pages 46–53, France, March 2000.

[64] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 33(12), 2011.

[65] Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M. Seitz. Being john malkovich. In *European Conference on Computer Vision*, 2010.

[66] Sander Koelstra and Maja Pantic. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. In *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September 2008.

[67] Irene Kotsia and Ioannis Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions On Image Processing*, 16(1):172–187, 2007.

[68] K. Kühnlenz, Stefan Sosnowski, and Martin Buss. The impact of animal-like features on emotion expression of robot head eddie. *Journal of Advanced Robotics*, 24(8-9), 2010.

[69] Fernando De la Torre, Joan Campoy, Zara Ambadar, and Jeff F. Conn. Temporal segmentation of facial behavior. In *Proceedings of the IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[70] K. M. Lam and H. Yan. Locating and extracting the eye in human face images. *Pattern Recognition*, 29(5):771–779, 1996.

[71] S.H. Leung, S.L. Wang, and W.H. Lau. Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Transactions on Image Processing*, 13(1):51–62, 2004.

[72] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *IEEE International Conference on Image Processing*, pages 900–903, 2002.

[73] Alan Wee-Chung Liew, Shu Hung Leung, and Wing Hong Lau. Segmentation of color lip images by spatial fuzzy clustering. In *IEEE Transactions on Fuzzy Systems*, 2003.

[74] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24:615–625, 2006.

[75] Arie A. Livshin and Xavier Rodet. The importance of cross database evaluation in musical instrument sound classification: A critical approach. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2003)*, Baltimore, Maryland, United States of America, 2003.

[76] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[77] Simon Lucey, Sridha Sridharan, and Vinod Chandran. Initialized eigenlip estimator for fast lip tracking using linear regression. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, page 3182, Los Alamitos, CA, USA, 2000. IEEE Computer Society.

[78] Mark G. Frank Claudia Lainscsek Ian R. Fasel Marian Stewart Bartlett, Gwen C. Littlewort and Javier R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6), 2006.

[79] Christian Martin, Uwe Werner, and Horst-Michael Gross. A real-time facial expression recognition system based on active appearance models using gray images and edge images. In *Proceedings of the International Conference on Face and Gesture Recognition*, 2008.

[80] David Matsumoto and Paul Ekman. American-japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion*, 13(2):143–157, June 1989.

[81] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135 – 164, November 2004.

[82] Christoph Mayer and Bernd Radig. Learning displacement experts from multi-band images for face model fitting. In *Proceedings of the International Conference on Advancements Computer-Human Interaction*, 2011.

[83] Christoph Mayer, Matthias Wimmer, Freek Stulp, Zahid Riaz, Anton Roth, Martin Eggers, and Bernd Radig. A real time system for model-based interpretation of the dynamics of facial expressions. In *Proceedings of the International Conference on Face and Gesture Recognition*, 2008.

[84] M.Beigzahed and M.Vafadoost. Detection of face and facial features in digital images and video frames. In *Proceedings of the 4th IEEE Cairo International Biomedical Engineering Conference*, 2008.

[85] A. Mehrabian. Communication without words. *Psychology Today*, 2:53–56, 1968.

[86] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[87] K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Audio- and Video-based Biometric Person Authentication, AVBPA'99*, pages 72–77, 1999.

[88] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 37(3), 2007.

[89] Louis-Philippe Morency and Trevor Darrell. Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 32–38, 2006.

[90] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2009.

[91] J. Nadel, M. Simon, P. Canet, R. Soussignan, Blancard, L. Canamero, and P. Gaussier. Human responses to an expressive robot. In *Proceedings of the sixth International Workshop on Epigenetic Robotics*, 2006.

[92] Daneil Neilson and Yee-Hong Yang. Evaluation of constructable match cost measures for stereo correspondence using cluster ranking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[93] Minh Hoai Nguyen and Fernando De la Torre Frade. Learning image alignment without local minima for face detection and tracking. In *8^{th} IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[94] T.D. Orazio, M. Leo, G. Cicirelli, and A. Distante. An algorithm for real time eye detection in face images. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.

[95] Mustafa Ozuysal, Vincent Lepetit, Francois Feuret, and Pascal Fua. Feature harvesting for tracking-by-detection. In *Proceedings of the European Conference on Computer Vision*, 2006.

[96] M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int'l Conf. Multmedia and Expo (ICME'05)*, 2005.

[97] Maja Pantic and Marian Stewart Bartlett. Face recognition. In *Machine Analysis of Facial Expressions*, pages 377–416, Vienna, Austria, 2007. I-Tech Education and Publishing.

[98] Maja Pantic and Ioannis Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on System, Man, Cybernetics*, 36(2), 2006.

[99] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[100] Sungsoo Park and Daijin Kim. Spontaneous facial expression classification with facial motion vectors. In *Proceedings of the 8^{th} IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September 2008.

[101] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

[102] S.L. Phung, A. Bouzerdoum, and D.Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1), 2005.

[103] Giacomo Rizzolatti Pier Francesco Ferrari, Vittorio Gallese and Leonardo Fogassi1. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience*, 17, 2003.

[104] Russel Reed. Pruning algorithms – a survey. *IEEE Transaction on Neural Networks*, 4(5):740–747, 1993.

[105] Lionel Revéret. From raw images of the lips to articulatory parameters: a viseme-based approach. In *In Proceedings of the 5th EuroSpeech Conference*, pages 2011–2014, Rhodos, Greece, 1997. University of Patras, Wire Communication Laboratory, Patras, Greece.

[106] Laurel D. Riek and Peter Robinson. Real-time empathy: Facial mimicry on a robot. In *Proceedings of the Workshop on Affective Interaction in Natural Environments*, 2008.

[107] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27, 2004.

[108] Sami Romdhani. *Face Image Analysis using a Multiple Feature Fitting Strategy*. PhD thesis, University of Basel, Computer Science Department, Basel, CH, January 2005.

[109] J.A. Russell and J.M. Fernandez-Dols. *The Psychology of Facial Expression*. Cambridge Univ. Press, 1997.

[110] Yunus Saatci and Christopher Town. Cascaded classification of gender and facial expression using active appearance models. In *International Conference on Face and Gesture Recognition*, 2006.

[111] M.T. Sadeghi, J.V. Kittler, and K. Messer. Modelling and segmentation of lip area in face images. *Vision, Image and Signal Processing*, 149(3):179–184, June 2002.

[112] Vahideh Sadat Sadeghi and Khashayar Yaghmaie. Vowel recognition using neural networks. *International Journal of Computer Science and Network Security*, 2006.

[113] Ashok Samal and Prasana A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25, 1992.

[114] Maja Pantic Sander Koelstra and Ioannis (Yiannis) Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32(11), 2010.

[115] Jason Saragih and Roland Gocke. Learning aam fitting through simulation. *Pattern Recognition*, 2009.

[116] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to AAM fitting. In *Proceedings of the International Conference on Computer Vision*, 2007.

[117] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Deformable model fitting with a mixture of local experts. In *Proceedings of the International Conference on Computer Vision*, 2010.

[118] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Face alignment through subspace constrained mean-shifts. In *Proceedings of the International Conference on Computer Vision*, 2010.

[119] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers, and T.S. Huang. Authentic facial expression analysis. *Image and Vision Computing*, Volume 25, Issue 12:1856–1863, December 2007.

[120] Mohammad Shami and Werner Verhelst. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49(3), 2007.

[121] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27:803–816, 2009.

[122] Caifeng Shan, Shaugang Gong, and P.W. McOwan. Robust facial expression recognition using local binary patterns. In *International Conference on Image Processing*, 2005.

[123] Yun Sheng, Abdul H. Sadka, and Ahmet M. Kondoz. Automatic single view-based 3-d face synthesis for unsupervised multimedia applications. 18, 2008.

[124] Gurvinder Singh Shergill, Abdolhossein Sarrafzadeh, Olaf Diegel, and Aruna Shekar. Computerized sales assistants: The application to measure consumer interest. *Journal of Electronic Commerce Research*, 2008.

[125] Nikhil V Shirahati and Kobus Barnard. Evaluating image retrival. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition*, 2007.

[126] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen. Skin detection in video under changing illumination conditions. In *15^{th} International Conference on Pattern Recognition*, pages 839–842, 2000.

[127] Stefan Sosnowksi, Christoph Mayer, Kolja Kühnlenz, and Bernd Radig. Mirror my emotions! combining facial expression analysis and synthesis on a robot. In *The 36th Annual Convention of the Society for the Study of Artificial Intelligence and Simulation Behaviour*, 2010.

[128] Mikkel Bille Stegmann and R. Larsen. Multi-band modelling of appearance. *Image and Vision Computing Journal*, 21(1):61–67, 2003.

[129] C. Stiller, G. Färber, and S. Kammel. Cooperative cognitive automobiles. In *Intelligent Vehicles Symposium, 2007 IEEE*, pages 215–220, June 2007.

[130] Simon Baker Terence Sim and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[131] A. Thayananthan, R. Navaratnam, B. Stenger, P. Torr, and R. Cipolla. Multivariate relevance vector machines for tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 124–138, 2006.

[132] Niels Henze Thomas Schlömer, Benjamin Poppinga and Susanne Boll. Gesture recognition with a wii controller. In *2nd International Conference on Tangible and Embedded Interaction*, 1999.

[133] Ying-Li Tian, Takeo Kanade, and Jeffrey F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[134] Marko Tscherepanow, Matthias Hillebrand, Frank Hegel, Britta Wrede, and Franz Kummert. Direct imitation of human facial expressions by a user-interface robot. In *Proceedings of the International Conference on Humanoid Robots*, 2009.

[135] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Graphics and Media Laboratory, Faculty of Computational Mathematics and Cybernetics*, Russia, 2003.

[136] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[137] P.J. Phillips W. Zhao, R. Chellappa and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35, 2003.

[138] Frank Wallhoff. The feedtum database. http://cotesys.mmk.e-technik.tu-muenchen.de/isg/content/feed-database, 2006. [Online; accessed 13-June-2011].

[139] Frank Wallhoff, Tobias Rehrl, Christoph Mayer, and Bernd Radig. Real-time face and gesture analysis for human-robot interaction. In *Proceedings of the SPIE, Society of Photo-Optical Instrumentation Engineers Conference*, 2010.

[140] Y. Wang and I. Witten. Inducing model trees for continuous classes. In *9th European Conference on Machine Learning*, pages 128–137, Prague, Czech Republic, April 1997.

[141] Paul Watzlawick, Janet Beavin Bavelas, and Donald D. Jackson. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. W. W. Norton and Co/NY, 1967.

[142] Wikipedia. Kinect. http://en.wikipedia.org/wiki/Kinect. [Online; accessed 18-March-2011].

[143] Wikipedia. Metcalfe's law. http://en.wikipedia.org/wiki/Metcalfe [Online; accessed 26-June-2011].

[144] Oliver Williams, Andrew Blake, and Roberto Cipolla. Sparse bayesian learning for efficient visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1292–1304, 2005.

[145] Matthias Wimmer, Freek Stulp, Sylvia Pietzsch, and Bernd Radig. Learning local objective functions for robust face model fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(8):1357–1370, 2008.

[146] Matthias Wimmer, Ursula Zucker, , and Bernd Radig. Human capabilities on video-based facial expression recognition. In *Proceedings of the Workshop on Emotion and Computing - Current Research and Future Impact*, 2007.

[147] H. Wu, X. Liu, and G. Doretto. Face alignment using boosted ranking models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[148] Rui Xiao, Qijun Zhao, David Zhang, and Pengfei Shi. Facial expression recognition on multiple manifolds. *Pattern Recognition*, 2010.

[149] Wenhui Liao Yan Tong and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *Transaction on Pattern Analysis and Machine Intelligence*, 29(10), 2007.

[150] Peng Yang, Qingshan Liu, and Dimitris N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *Proceedings of the twelfth International Conference on Computer Vision*, September 2009.

[151] Jeffrey F. Cohn Yang Wang, Simon Lucey. Enforcing convexity for improved alignment with constrained local models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[152] Jeffrey F. Cohn Yang Wang, Simon Lucey. A generative shape regularization model for robust face alignment. In *Proceedings of the European Conference on Computer Vision*, 2008.

[153] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3D dynamic facial expression database. In *Proceedings of the 8^{th} IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September 2008.

[154] Hui Zhang, Sharath Cholleti, and Sally A. Goldman. Meta-evaluation of image segmentation using machine learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2006.

[155] Glenn I. Roisman Zhihong Zeng, Maja Pantic and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2009.

[156] Jianke Zu, Luc Van Gool, and Steven C.H. Hoi. Unsupervised face alignment by robust nonrigid mapping. In *Proceedings of the International Conference on Computer Vision*, 2009.