Facial Communicative Signals: Valence Recognition in Task-Oriented Human-Robot Interaction

Christian Lang

Dipl.-Inform. Christian Lang
CoR-Lab / Angewandte Informatik
Technische Fakultät
Universität Bielefeld
email: christian.lang@uni-bielefeld.de

Abdruck der eingereichten Dissertation zur Erlangung
des akademischen Grades Doktor-Ingenieur (Dr.-Ing.).
Der Technischen Fakultät der Universität Bielefeld
am 27.06.2012 vorgelegt von Christian Lang.

Betreuer:

- PD Dr.-Ing. Sven Wachsmuth
- Dr.-Ing. Marc Hanheide
- Dr. rer. nat Heiko Wersing

Gutachter:

- PD Dr.-Ing. Sven Wachsmuth
- Dr.-Ing. Marc Hanheide
- Assoc. Prof. Dr. Modesto Castrillón-Santana

Prüfungsausschuss:

- Prof. Dr. rer. nat. Barbara Hammer
- PD Dr.-Ing. Sven Wachsmuth
- Dr.-Ing. Marc Hanheide
- Assoc. Prof. Dr. Modesto Castrillón-Santana
- Dr. rer. nat. Thies Pfeiffer

Gedruckt auf alterungsbeständigem Papier nach ISO 9706

Facial Communicative Signals: Valence Recognition in Task-Oriented Human-Robot Interaction

Dissertation zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

der Technischen Fakultät der Universität Bielefeld

vorgelegt von

Christian Lang

Bielefeld – Juni 2012

Acknowledgments

Writing a dissertation is hard work. I am very grateful that I received a lot of help from many people. First of all, I thank my supervisors Sven Wachsmuth, Marc Hanheide, and Heiko Wersing for their great support during the whole project, and also Modesto Castrillón-Santana for being a reviewer of my thesis. I am grateful to Frank Hegel, Werner Schneider, Gernot Horstmann, and Angelika Dierker for the useful discussions about some aspects of this work. A great thank-you goes to Matthias Schöpfer for his superb management of the citec compute cluster I used rather frequently in the recent months. I thank Sascha Hinte, Anton Helwart, and Benjamin Koch for their help in conducting studies and annotating videos, and I am also very thankful to all the people who participated in the conducted studies. I want to thank the members of the Applied Informatics group and the CoR-Lab in general for the very kind working atmosphere.

I am very grateful to my wife Annelie for her love and immense support and also to my son Noah Daniel for giving me so much joy. I also want to thank my parents and parents in law for all their help. The greatest thank of all goes to God for his love, great support, and motivation throughout my PhD project.

Abstract

In this dissertation, we investigate facial communicative signals (FCSs) in terms of valence recognition in task-oriented human-robot interaction. Facial communicative signals mainly comprise head gestures, eye gaze, and facial expressions. We review important psychological findings about the human display and perception of FCSs. Based on this discussion, several conclusions are drawn that motivate the presented work.

We investigate a FCS recognition in terms of positive or negative valence in an object-teaching scenario where human subjects teach objects to a robot. The correct or wrong answer of the robot when queried for the object name is used to define the ground truth data for the FCSs the humans displayed in turn during their reaction to this answer. Thus, the facial display the human showed after the robot classified an object correctly is treated as an example of the positive or *success* class. Similarly, the FCSs shown after a wrong answer constitutes an example of the negative or *failure* class. We evaluated to which degree humans can infer whether the answer of the robot was correct or not from looking at these facial displays only.

Furthermore, we present a simple static baseline approach for the automatic classification of these facial displays in terms of valence. It is based on feature extraction with active appearance models (AAMs) and a classification with support vector machines (SVMs). The method does not consider temporal dynamics, but uses a simple majority voting scheme over the classification results for the single frames.

This simple static approach yielded baseline results for a more sophisticated dynamic approach. The dynamic approach is based on the selection of discriminative reference subsequences as prototypes in a nearest-neighbor-based classification scheme. The temporal dynamics are considered by means of dynamic time warping (DTW) which is used to compare sequences of AAM feature vectors. In the conducted evaluation, this dynamic FCS recognition approach outperformed the static baseline approach and achieved human level classification accuracies in a person-specific classification.

Contents

1. Introduction

There is no such thing as "natural" human-robot interaction.
Human-robot interaction is always artificial.
— Kerstin Dautenhahn

Human-robot interaction is a very exciting and rapidly developing interdisciplinary research field. It relies on knowledge and methods from psychology, sociology, engineering, and informatics and has made impressive progress in recent years. One overarching goal of research in this area is to make interactions with robots more and more user-friendly. An encounter with a robot should be enjoyable and uncomplicated. Even though the interaction with a machine might never become truly natural in the sense as human-human interactions are—at least not if Prof. Dautenhahn [104] is right—it can become increasingly effective: the robot shall adapt to the human, not vice versa.

A crucial aspect in order to come closer to this goal is the way of communication. Besides speech, also nonverbal channels play a vital role in this respect [515]. Thus, equipping a robot with sufficient nonverbal communication skills is an important step towards the desired effectiveness of interaction. This dissertation contributes to this goal by investigating one essential way of nonverbal communication: facial communicative signals.

1.1. Scope and Contribution

The scope of this work is the investigation of facial communicative signals in human-robot interaction scenarios. The general research questions are:

How can a robot be enabled to effectively recognize and interpret the facial displays of its human interaction partner?

How can the robot use the (implicit) feedback signals given by the human to make the interaction more efficient?

This dissertation contributes to the answering of these questions.

1.1.1. Facial Communicative Signals

Facial Communicative Signals (FCSs) do generally comprise head gestures, eye gaze, and facial expressions. Regarding their exact definition, we will take a pragmatic view that focuses on the information about the interaction that can be gained from observing facial displays. This is motivated by a review of major psychological findings concerning FCSs, which shows the immense complexity of this aspect of human behavior.

Parts of this dissertation have been published before [304, 305, 308, 306, 307]. Christian Lang is the first author and writer of all these publications. Please refer to App. A.8 for details.

Figure 1.1.: Three robots that have been used in various human-robot interaction studies at Bielefeld University. Left: The iCub robot platform [239]. Middle: Biron, the *BIelefeld Robot CompaniON* [204]. Right: The Nao robot [420].[1]

1.1.2. Valence Recognition

We will argue that because of this complexity, a pragmatic simplification is necessary to make progress in this area of human-robot interaction. The approach we suggest is an interpretation of FCSs in terms of positive or negative *valence*, because this has several significant advantages compared to other approaches. Applied to human-robot interaction, we will relate this valence to successful or problematic interactions with the robot.

1.1.3. Task-Oriented Human-Robot Interaction

We will further argue that an investigation of FCSs "in general", without being related to a certain context, is not feasible in practice. Therefore, FCS investigations need to be conducted within specific interaction scenarios that imply certain contexts. We will focus on *task-oriented* interaction scenarios where the robot is to perform a certain task, possibly in collaboration with its human interaction partner. As one concrete example, we will investigate a scenario where our robot Biron (please see Fig. 1.1) shall learn objects that are taught be the human.

1.2. Challenges

Before we outline the structure of the thesis, we want to give some anecdotal impressions of the related challenges and open questions by quoting a few researchers of this field. Regarding the questions "How can we find out what people are thinking and feeling?" and "How do people express this?", Heylen *et al.* [218, p. 6] wrote in their conclusion:

[1] Photographer of the iCub image: Barbara Proschak

> "When one starts collecting and analysing naturalistic data, it becomes immediately clear that one cannot rely on the standard associations between behaviours and functions that the existing research has focused on (for instance the relation between facial expressions and emotions). Careful collection and analysis of data should enable us to construct a more accurate picture of the associations between expressions and meanings."

Similarly, regarding the concept of *basic emotions* [133], Zeng *et al.* [548, p. 41] state in their recent survey of affect recognition methods:

> "However, discrete lists of emotions fail to describe the range of emotions that occur in natural communication settings. For example, although prototypical emotions are key points of emotion reference, they cover a rather small part of our daily emotional displays. Selection of affect categories that can describe the wide variety of affective displays that people show in daily interpersonal interactions needs to be done in a pragmatic and context-dependent manner."

Murphy-Chutorian and Trivedi [368, p. 607] begin their recent survey of head pose estimation in computer vision with this assessment:

> "From an early age, people display the ability to quickly and effortlessly interpret the orientation and movement of a human head, thereby allowing one to infer the intentions of others who are nearby and to comprehend an important nonverbal form of communication. The ease with which one accomplishes this task belies the difficulty of a problem that has challenged computational systems for decades."

In their paper titled "Searching for Prototypical Facial Feedback Signals", Heylen *et al.* [217, p. 148] summarize some properties of human communicative behaviors:

> "Some important characteristics of expressive communicative behaviours are that (a) a behaviour can signal more than one function at the same time, (b) behaviours may serve different functions depending on the context, (c) and behaviours are often complexes composed of a number of behaviours. Moreover, (d) the absence of some behaviour can also be very meaningful."

1.3. Thesis Structure

The next chapter reviews some important psychological findings about the human display and perception of facial communicative signals, in particular head gestures, eye gaze, and facial expressions. We draw several conclusions from this discussion that motivate the approach of valence recognition in an object-teaching scenario that is presented in the third chapter. This chapter also investigates the human recognition performance in this scenario.

The fourth chapter gives an overview of state of the art approaches for an automatic recognition of FCSs. In the fifth chapter, we consider a simple static approach for the automatic classification of FCSs in our object-teaching scenario. This approach operates on single frames without considering temporal dynamics and uses support vector machines to perform the classification.

The results of this static method provide a baseline for the performance of a more sophisticated, dynamic recognition approach that is investigated in the sixth chapter. It is based

on the selection of discriminative reference subsequences as prototypes for the classification and considers the temporal dynamics by applying dynamic time warping as distance measure. This approach outperforms the static baseline approach and achieves human level classification accuracies in a person-specific FCS classification. Finally, the last chapter concludes the thesis and comments on future research.

2. Facial Communicative Signals

One cannot not communicate.
— Paul Watzlawick

Paul Watzlawick conducted widely recognized research on human communication. Together with Janet B. Bavelas and Don D. Jackson, he published the book *Pragmatics of Human Communication* [515], in which they presented an interpersonal communication theory consisting of five axioms. The first axiom is known as "one cannot not communicate", stating that every behavior has a communicative aspect. As there is no opposite or alternative to "behavior", none is to communication as well. This especially emphasizes the importance of nonverbal communication in social interactions, as a communicative meaning might be attributed to every nonverbal action, whether intended or not.

The communicative meaning of nonverbal behaviors emerges from a complex interaction of face and body and is also significantly influenced by the environment and context in many respects. We focus our investigations on the face, because it is one of the richest means of nonverbal communication and plays a crucial role in social interactions by itself, notwithstanding the high importance of other nonverbal behaviors such as hand or body gestures. However, these different facets of nonverbal communication are not disjunct. For instance, head gestures concern not only the face, but the body as well.

Some aspects of human face perception in general are briefly discussed in Sec. 2.1. Subsequently, facial communicative signals, especially head gestures, eye gaze, and facial expressions, are discussed in greater detail in Sec. 2.2. Finally, Sec. 2.3 draws conclusions from this discussion and gives a first motivation for our valence-based approach to facial communicative signal recognition, which is explained in the following chapter.

2.1. Human Face Perception

This section briefly introduces some major findings about the human face perception in general. The physiological basis is discussed in Sec. 2.1.1, followed by a consideration of several important face perception effects found in behavioral studies in Sec. 2.1.2. Subsequently, some face perception features (Sec. 2.1.3) and also a few developmental aspects (Sec. 2.1.4) are considered, before Sec. 2.1.5 concludes this discussion of the human face perception.

2.1.1. Physiological Basis

Haxby *et al.* [211] presented a hierarchical model of the human neural face perception system (Fig. 2.1). The main components are a *core system* that represents invariant (identity) and

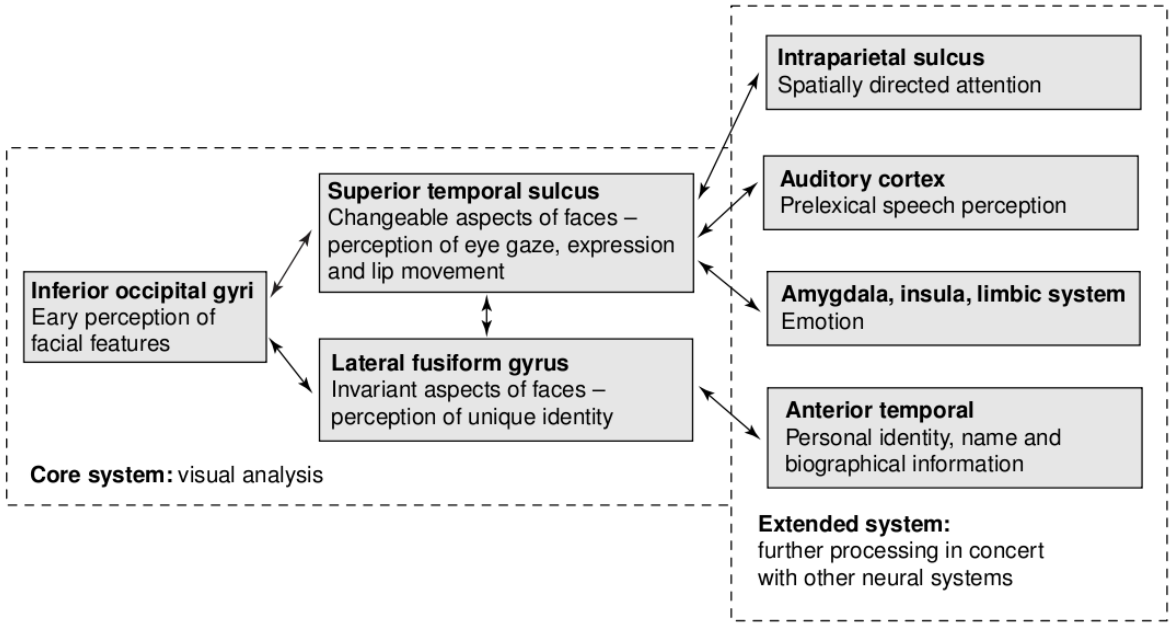


Figure 2.1.: Summary of the human neural face processing model presented by Haxby *et al.* [211]. The model proposes a core system for the visual analysis of invariant and changeable aspects of face stimuli and an extended system for the inference of further attributes. Please refer to Sec. 2.1.1. Reproduced with kind permission.[2]

changeable (expression, eye gaze) aspects of faces in different, dedicated brain regions, and an *extended system* which recruits other functional brain areas for further interpretations (e.g. spatial attention, emotion, and name). The model proposes a distributed processing where many face perception functions rely on the interaction of several brain regions. It refines and extends an earlier model by Bruce and Young [49].

Several investigations provide evidence for such a model. Studies with patients suffering from prosopagnosia[1] [156] suggest the existence of a specialized face perception system. Various neuroimaging studies using *positron emission tomography* (PET) [497] (notably the studies by Sergent *et al.* [453] and Haxby *et al.* [212]), or more recently *functional magnetic resonance imaging* (fMRI) [235] (especially the studies by Clark *et al.* [81], Kanwisher *et al.* [270] and McCarthy *et al.* [350]) show high activations in specific brain areas when the participants looked at faces. In particular, Kanwisher *et al.* [270] conducted a series of experiments showing that in the *fusiform face area* (FFA) in the *right fusiform gyrus* the activation was significantly higher for faces than for other objects for most tested subjects. Furthermore, they ruled out several alternative explanations (low-level feature extraction, visual attention, response to human or animate objects in general [271], recognition of objects of the same category or luminance) and concluded that the FFA is selectively activated by various kinds of face stimuli. While the FFA seems to be essential for face identification, another brain area in the *superior temporal sulcus* (STS) appears to be more important for eye gaze, facial expression and dynamic facial movement processing [461]. Calder and Young [59] presented a critical review of the evidence supporting such a separation.

[1] Prosopagnosia is a face perception disorder impairing the patient's ability to identify faces, often even very familiar ones, while the capability the recognize objects is largely intact.

However, this specialization for face processing has been questioned in subsequent research. Gauthier *et al.* [181] demonstrated that expertise with cars and birds can yield high activations in the FFA and concluded that this brain region is rather related to level of expertise than to visual appearance of objects or faces. Similar results for cars were obtained by Xu [529] and also for "greebles"[3] by Gauthier *et al.* [184] in earlier work. Rossion *et al.* [421] presented some evidence that non-face objects of visual expertise can compete with faces for early neural processing. Hanson and Halchenko [208] criticized the common view that high activations indicate specialization and suggested experiments with statistical classifiers to find specialized brain regions instead. They performed classifications with *support vector machines* (SVMs) [100] using neuroimages of the whole brain to distinguish faces from houses and found a distributed code but no single areas that were discriminative for faces or houses. Haxby *et al.* [210] also proposed a distributed and overlapping representation of faces and objects and concluded from their experimental results that besides the regions of maximal activation, also areas of smaller responses contribute vitally to the representation.

Kanwisher and her colleagues were not convinced of most of the criticism nevertheless [268]. Based on a comprehensive review of the research [272], they defended the face specificity of the FFA and rejected the alternative explanations arguing for utilized general-purpose processes, for instance the expertise hypothesis. Moreover, she interpreted the study of Tsao *et al.* [490], who found a very high face selectivity in a specific brain region in macaque monkeys using single cell recording, as providing very strong evidence for face-specific brain areas also in humans [269]. This view is also supported by Tsao and Livingstone [491], who additionally suggest the existence of a specific detection process that makes faces special. Altogether, even though it is clear that the FFA and some other brain areas play important roles, the human neural face perception system still poses various open questions and its comprehensive understanding is subject to further research.

2.1.2. Behavioral Studies

Besides neuroimaging studies, a large body of behavioral studies has been conducted to determine the basic properties of the human face perception. There is evidence that normal, upright faces are processed in a *configural* or *holistic* way [476, 477], in contrast to *parts-based* representations of objects [478, 494], although the exact definition of *configural* and *parts-based* processing is controversial to some extent [418]. As discussed below, several standard tasks showing certain effects have been associated with configural processing. It is commonly accepted that these effects occur for faces, but not or to a significantly lesser extent for objects in general (e.g. [184, 418, 157]), but whether they emerge for objects of expertize is controversial again. While Gauthier *et al.* [184] referred to studies demonstrating similar behavioral effects for experts of non-face objects [113, 50, 182], Robbins and McKone [418] denied this based on a research review and own studies with dog experts who viewed images of their breed-of-expertise:

- *Inversion effect*:[4] While the expected strong inversion effect for faces occurred [542],

[2]Reproduced from the original paper [211, Fig. 5, p. 230] with kind permission of Elsevier and James V. Haxby. Copyright © 2000 Elsevier Science Ltd.

[3]The "greebles" used in this study were computer-rendered images of artificial, virtual objects that roughly resemble the T-shaped configuration of human faces. They are divided in different "families", consisting of certain "individuals".

[4]People recognizing faces from images perform significantly better when the face images are presented upright

no strong inversion effect was found for dog images, which is in contrast to the earlier experiments of Diamond and Carey [113] who found such an effect. The authors consider this in line with other results (e.g. [50, 530]) and point out that Diamond's and Carey's results [113] have not been replicated so far.

- *Composite effect*:[5] A clear composite effect for upright faces, but not for inverted faces, was found as expected (e.g. [545, 223]). For dog images no such effect was observed, neither for novices nor experts. Robbins and McKone [418] view this as a confirmation and extension of Gauthier and Tarr's results for "greebles" [183], while Tarr [480] interprets the latter differently as showing a composite effect for "greebles".
- *part-whole effect*:[6] Several studies demonstrated a strong part-whole effect for upright faces, but no or a smaller effect for inverted faces and objects [476, 182, 105]. It is controversial to which degree the part-whole effect indicates holistic processing. Gauthier and Tarr [183] argue that it assesses some generic context advantage only, while Robbins and McKone [418] suggest that this generic context effect is rather small compared to the "true" (face-specific) effect arising from perceptual integration. Donnelly and Davidoff [117] found a strong part-whole effect also for images of houses, but rejected holistic processing as explanation in subsequent experiments. Generally, the composite effect is considered a better indicator for holistic processing.
- *Contrast reversal effect*:[7] The experiments of Robbins and McKone [418] showed a strong contrast reversal effect for faces as expected [48, 178], but only a small one or none for dogs, similar to the results of Gauthier *et al.* [185].[8] Although the contrast reversal effect is not directly related to configural processing and can occur independent of other effects [48, 259], it is another potentially face-specific effect.

In reaction to the argumentation of Robbins and McKone [418], Gauthier and Bukach [180] criticized Robbins and McKone's treatment of the composite effect and argued that also small behavioral effects are important as the higher magnitude of these effects in faces might arise from the outstanding, life-long experience with faces, so expertize might well be the cause of these effects, for which they see strong evidence. Moreover, Bukach *et al.* [54] pointed out the general value of the expertize framework as a research tool also beyond face perception. McKone and Robbins [355] countered this criticism on several levels and suggest to reject the expertize hypothesis as far as faces are concerned.

Besides the issues discussed above, also other face processing effects have been considered. Schweinberger *et al.* [449, 448] investigated the relations between the perception of face identity and facial expressions and reported an asymmetric interaction: the processing of facial expressions seems to depend on the perceived identity, but not vice versa. This contrasts previous findings of Bruce [45] who did not find a similar interaction and argued for an independence of identify and facial expression perception. Recently, Soto and Wasserman [468] presented evidence that pigeons show a similar asymmetrical interaction. As pigeons

than when presented upside down ("inverted").

[5] When an upper half and a lower half of two different face images are combined to form a new image, people can recognize either half more easily when the two halves are horizontally displaced than when they are aligned (and thus form a "new" face).

[6] People can identify face parts (e.g. John's mouth) better when they are presented as part of a whole face, compared to isolated presentation.

[7] Contrast reversal of face images impairs the recognition performance of viewing people.

[8] Interestingly, Robbins and McKone [418] found no contrast reversal effect for dog experts, but a small effect for novices, while Gauthier *et al.* [185] found a small effect for "greeble" experts and none for novices.

are unlikely to feature a specialized system for the perception of human faces, this suggests a generic processing mechanism as the source of these effects. Thus, the debate about the specialness of face processing and the associated effects is far from being over.

2.1.3. Face Processing Features

Humans can identify familiar faces at image resolutions as low as 7 × 10 pixel well above change level, where ceiling level is reached at approximately 25 × 30 pixel [461]. A distinction between faces and non-face objects is possible at an even lower resolution of 7 × 7 pixel [31]. Several studies showed that people can recognize familiar faces better than unfamiliar ones [47, 55, 417]. Besides the actual face area with inner features such as eyes, nose and mouth, humans appear to make strong use of the overall head shape to identify people, as a famous experiment of Sinha and Poggio [462] with photographs of Bill Clinton and Al Gore manipulated to show identical inner faces demonstrate.

The human recognition performance depends on the illumination of the face, but this effect is small for familiar faces under usual circumstances. Generally, illumination-induced changes appear to be included in the facial representation [461]. Evidence suggests that both face shape and texture are important cues to perform face recognition [461]. By evaluating several studies, Bruce *et al.* [46, p. 293] concluded that certain *surface-properties* seem to be very important for the human visual system. They further reviewed frequency-decomposition experiments which demonstrated the relatively high relevance of low-level features, compared to object recognition [46, p. 289]. The holistic (no decomposition into parts) or at least configural (face parts are not perceived independently of each other) processing of faces has already been discussed in Sec. 2.1.2. Some face processing tasks can be performed very fast, so they probably require only one feed-forward pass through the visual system [461].

2.1.4. Developmental Aspects

Even very young neonates show a probably innate preference for "face-like" patterns (e.g. [257]), although this is controversially debated (e.g. [495]). Possible consequences for the neural development [256] and also the acquisition of adult face expertize [179] have been discussed. Newborns can distinguish their mothers from strangers and might already be able to imitate some facial expressions, but the last finding is controversial [461]. Within a face, two-month-old infants are especially looking at the eyes [345]. It has been shown that three-month-old babies can shift their visual attention in the direction indicated by an adult face [225], and at the age of four months, they can discriminate direct from averted gaze [501]. Further face recognition capabilities develop gradually until approximately ten years of age [461]. This includes an increasingly configural encoding of faces [243, 61]. For a deeper discussion, please refer to the book of Johnson and Morton [258]. They presented a theory of the development of face recognition that proposes two basic components: *conspec*, which refers to innate structural information about faces, and *conlern*, which concerns various learning mechanisms that lead to sophisticated face recognition capabilities.

2.1.5. Conclusion

Many studies have investigated the traits of human face perception, in part with disputed results, as discussed above. Nevertheless, we can state some important conclusions to keep in

mind for the following investigations:

- The human brain appears to process invariant (identify) and changeable (facial expressions, eye gaze, etc.) aspects of faces by different neural subsystems. However, this does not mean that the final perception and interpretation of these aspects is also independent. In particular, the perceived face identify appears to have an influence on the perception of facial expressions.
- Adult humans process faces in a *holistic* way, in contrast to a *parts-based* processing of (most) other objects.
- Humans are considerably better in recognizing familiar faces than in recognizing unfamiliar ones, especially in low-resolution images.
- Both the shape and the texture of a face appear to be important cues for the human face perception.
- Appearance variations due to different illuminations appear to be included in the face representation.

Please refer to the review of Sinha *et al.* [460] for a deeper discussion about the consequences the findings about the human face perception have for computer vision research.

2.2. Facial Communicative Signals

Our definition of *facial communicative signals* does not focus on the visual appearance of the face. Instead, it emphasizes the attribution of meaning to a certain facial behavior:

Facial Communicative Signal (FCS): Any visual facial behavior in interaction situations that can be meaningfully interpreted as nonverbal feedback and can thus be utilized to infer some relevant information about the course of interaction resp. the interaction partner who shows this facial behavior.

This is a pragmatic view. It does not distinguish between deliberately given *signals* and unintentionally present *cues* [348], but focuses on their meaningful interpretability in an interaction situation only. Furthermore, it does neither define which facial behaviors exactly are FCSs and which are not, nor does it specify what exactly an attributable "relevant" meaning is. In fact, both issues usually depend on the context of the interaction (please refer to Sec. 2.3 and Sec. 3.7). Similarly to our view, DeCarlo *et al.* [109] used the term *facial conversational signals* to describe signals of this kind. They introduced them informally by examples, but did not state a definition. Cerrato and Skhiri [67] considered *human communicative gestures*, which also include head movements, eye gaze, and facial expressions, among others such as hand gestures and body posture.

Two examples of FCSs are depicted in Fig. 2.2. The interaction situation is quite similar in both cases: a robot did not perform an action as expected by the user. As a result of this, the user probably thinks about how to proceed with the interaction. However, the shown FCSs are very different in terms of visual appearance. Whether they are also different with respect to the attributed meaning depends on the intended kind of interpretation. For a comparatively precise interpretation, they can be viewed as beeing different, signalling "being puzzled" (left image) and "being cogitating" (or "looking down" in a more descriptive interpretation, right image), for instance. In case of a coarser interpretation, both FCSs might as well be regarded as having the same meaning, for instance "thinking about how to proceed" or "signalling some kind of failure".

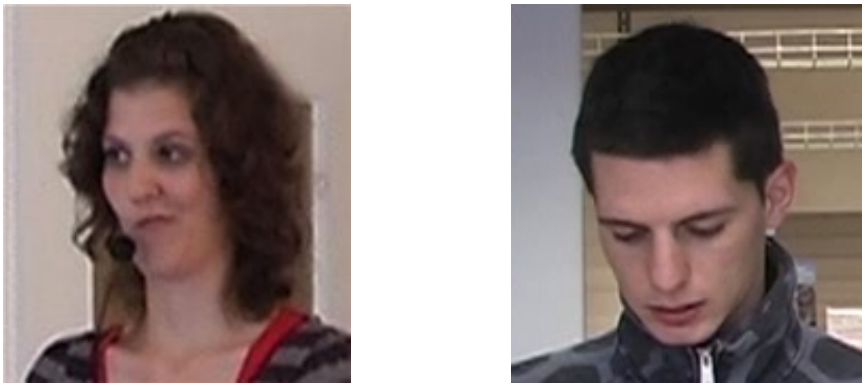


Figure 2.2.: Two examples of facial communicative signals (FCSs), both displayed after an interaction with a robot did not yield the desired result. Reasonable interpretations are "being puzzled" (left) and "being cogitating" or just "looking down" (right), but also "signalling some kind of failure" in both cases. Please refer to Sec. 2.2.

Despite the large variations due to different interaction contexts and desired kinds of interpretation, FCSs generally include head gestures, eye gaze, and facial expressions. These FCSs are discussed in the following sections in some detail.

2.2.1. Head Gestures

Despite their overt presence in conversations in everyday life and scientific investigations from several disciplines, head gestures are not nearly as extensively researched as facial expressions and other gestures [214]. Birdwhistell [33] distinguished several head movements and positions that constitute meaningful elements in conversations, including the following (according to Heylen [214]):

- full head nod (up and down or down and up)
- half head nod (either up or down)
- small "bounce" at the end of a (full or half) head nod
- full head sweep (to left or right and back)
- cocked head

He also differentiated a single head nod from two consecutive nods and three or more nods in a conducted study, where he looked into the effect these nods have on a speaker when given by an auditor (e.g. affirming or encouraging the speaker or make her resp. him hesitate). Several other researchers investigated head gestures and their communicative functions as well. Tab. 2.2.1 lists certain important head gestures together with their ascribed meaning. DeCarlo *et al.* [109] developed a virtual agent that uses head gestures (besides other signals) to enhance its verbal conversation with a user. Cerrato and Skhiri [67] analyzed the head movements of subjects acting as "information giver" during a conversation with an "information seeker". They concluded that head gestures have different meanings in different contexts. Poggi *et al.* [403] investigated in detail the meaning of head nods in political debates broadcasted on TV and also pointed out that their meaning heavily depends on the context and the role of the person who nods (speaker, interlocutor, listener); head nods in conversations between japanese people were analyzed by Maynard [347]. Iwano *et al.* [245] investigated head movements

head gesture	**intended meaning - DeCarlo *et al.* [109]**
nod downward	general indicator of emphasis
nod upward	perhaps indicate a "wider perspective"
bring whole head forward	perhaps need for "a closer look"
bring whole head backward	perhaps being "taken aback"
turn left or right	perhaps indicate availability of more information
tilt head	perhaps indicate expectated user engagement
tilt head + nod downward	perhaps indicate contrast of related topics

head gesture	**attributed meaning - Cerrato & Skhiri [67]**
nod (forward movement)	intent to continue ("You go on", "I want to go on"), acknowledgement, make a statement
jerk (backward movement)	acknowledgement, surprise
shake (move to left/right and back)	negative answer or statement, refusal
waggle (back and forth, left to right)	disfluencies in conversation

head gesture	**attributed meaning - Poggi *et al.* [403]**
nod	confirmation, agreement, approval, submission, permission, greeting, thanks, backchanneling and its request, emphasis, ironic agreement, literal and rhetoric question

head gesture	**attributed meaning - Maynard [347]**
nod	backchanneling, turn-taking, turn-transition filler, affirmation, agreement, emphasis

head gesture	**attributed meaning - Iwano *et al.* [245]**
vertical head movement	recognition success, content affirmation, and also their degree; agreement
horizontal head movement	content denial, but not degree of it
Inclined head movement	possibly withholding or scepticism
face up to interaction partner	response request

head gesture	**attributed meaning - McClave [351]**
nod	backchanneling and its request
lateral sweep	signal inclusivity (e.g. "everything", "whole"), intensification (e.g. "a lot", "exactly"), uncertainty (e.g. "I guess", "whatever"), lexical repairs
change in head orientation	mark direct quote of someones utterance, mental imagination, spatial referencing, talking about lists or alternatives

head gesture	**attributed meaning - Kendon [278, 279]**
change in head orientation	indicate start of new "speech unit" (e.g. sentence)
head shake	negation, denial, universal statement, intensification, self-commenting (e.g. self-correction)

Table 2.1.: Several head gestures and their functions as investigated by various researchers. Please refer to Sec. 2.2.1.

during spoken dialog and during cooperative crossword puzzle solving. They observed many head movements during listening and less during speaking turns, which is in contrast to the results of Hadar *et al.* [205], who found the converse pattern. From the cooperative crossword puzzle solving task, Iwano *et al.* [245] concluded that speech is sufficient when the interaction goes well, but when problems occur, head gestures become more important. Bavelas and Chovil [27] investigated the relation between speech and nonverbal acts, including head gestures, in detail. Some of their most important conclusions are that several nonverbal acts

- are fully integrated with the accompanying words in conversations
- may or may not be redundant to speech
- are "analogically encoded symbols" whose meaning must be assesed in the respective context

McClave [351] also explored the functions of head movements accompanying speech. She pointed out that while head nods and shakes are understood contextless as signalling "yes" resp. "no" (in american culture), several others need to be considered in their respective context, especially speech. Furthermore, she expected that some head gestures and their associated functions are culure-specific, whereas others (e.g. spatial referencing) might be universal and stressed the need for cross-cultural studies. Kendon analyzed the body language during a casual conversation [278] and also elaborated on the different functions of head shakes [279]. Graf *et al.* [195] also addressed the relation between speech and head gestures and found two different types of nods and one type of swing as typical movement patterns for a "visual prosody", despite large variations depending on several factors (e.g. personality and content of conversation). Several further investigations were undertaken, for instance a detailed analysis of "thinking faces" when people search for a word [191].

However, due to the complex interrelations of speech, head gestures, kind of interaction and its context, culture and personality of the involved people, most of these results need to be taken with care regarding their universality, especially considering the often small number of subjects in the conducted studies. Birdwhistell [33] emphasized that he focused his analysis on american people where he observed regional differences similar to dialects. He further stated that tentive, preliminary research concerning french, german, and english people suggests large cultural variations. Prominent examples of intercultural variations are also the differences between the indian "head wiggle" and western head gestures of comparable meaning [472], or the reversed meaning of head nods and shakes in Albania and Bulgaria, compared to most western countries [123]. DeCarlo *et al.* [109] clearly state the tentative character of their head gesture model, which was developed based on preliminary video analysis and "informal observations of everyday conversation" [109, p. 33]; they call the functions listed in the respective block of Tab. 2.2.1 "some rough speculations about the functions that these different movements might carry" [109, p. 33]. Cerrato and Skhiri's [67] results are based on recordings of two participants only, McClave [351] analyzed the behavior of four subjects, Kendon [278] did this for one person. Iwano *et al.* [245] analyzed the head movements of one japanese person only and regarded their results as being culturally bounded due to differences in nodding and answering behavior between japanese and english-speaking people (e.g. [347]). The four subjects in the study of Hadar *et al.* [205] were aware of the interest in their head movements during their recorded conversations, which might have had an influence on their behavior. (The same might be true for other studies, e.g. [67].)

Despite these limitations, the discussed studies evidently show that head gestures serve many different functions. Heylen [214] compiled a long list of these functions by reviewing the literature. Besides the already mentioned meanings, this list also includes (among others):

- enhancing communicative attention
- control and organization of the interaction
- marking the contrast with preceding utterances
- indicating interest

The concrete meaning that can be reansonably attributed to certain head gestures also depends on various aspects of the interaction context, most notably:

- the general situation (e.g. discussion, interview, casual conversation, cooperative task)
- the culture, socialization and personality of the involved people
- the accompanying speech
- the role of the person (e.g. speaker, interlocutor, listener)

As many researchers pointed out (e.g. [279]), additional research is required to further clarify the exact role of such contextual aspects and their complex interplay.

2.2.2. Eye Gaze

In contrast to other primates, the visual appearance of human eyes allows for advanced gaze-signalling and thus enhance communication [291], although other species make some use of gaze in social interactions [145], too. Eye gaze is a very special FCS, because it is both a channel to perceive visual information and a signal in social interactions at the same time [333]. People can estimate the gaze direction of others reasonably well, especially when being looked at (e.g. [82, 7]), where both eye gaze and head position mutually influence the perceived gaze direction [309]. The human brain seems to feature an expert system for gaze perception [416].[9] Baron-Cohen regards the detection of the eye gaze direction as an important component in his model of a *mindreading system* in the human brain [19].

The following section outlines the relationship of eye gaze and attention, before the role of gaze in social interactions is discussed afterwards. Finally, the concluding section summarizes the multiple functions eye gaze can serve.

Gaze and Attention

The apparently most important single aspect of eye gaze is to signal visual attention [312], it is also important for establishing joint attention [46, 132, 19, 399, 57]. According to Argyle and Cook [9], eye gaze is generally closely tied to attention in the perception of people, looking more is usually perceived as more attentive. Argyle and Graham [11] investigated the gazing behavior of people performing a task together in dyads. In case an object relevant to the task was present, it attracted the majority of gaze (even more if it was a complex object), while the persons looked at each other most of the time when no such object was there. This behavior could not be fully explained by the possible information gain resulting from looking

[9] This was concluded by Ricciardelli *et al.* [416] who showed in a series of experiments that contrast reversal of images showing human eyes highly impairs the gaze perception of observers. Thus, the human brain seems not to use purely geometrical information as proposed previously (e.g. [7, 312]).

at the object. The effect of task-irrelevant background stimuli on gaze was unreliable. Hugot [236] also reported a strong focus of eye gaze on a task-relevant object.

The automatic shift of visual attention to "cued" locations was researched in several studies. Langton and Bruce [310] evaluated the reaction time of people responding to target stimuli on a screen. They found that the display of a face gazing in the direction of the next target improved the reaction time, despite the subjects knew that this cue was not predictive and were told to ignore it. Very similar results were obtained by Friesen and Kingstone [172]. Using arrowheads instead of faces seem not to produce such an effect [262],[10] neither does the abrupt onset of a stimulus itself [173]. Taken together, these findings suggest that someones gaze direction can reflexively shift attention away to where this person is looking. Furthermore, there is some evidence that direct gaze of others can automatically draw attention to them [507]. However, Cooper [85] showed that these automatic shifts of attention do not occur in any case: In several experiments, he found no evidence that direct gaze was *particularly* difficult to ignore, at least no more difficult than a face with closed eyes. He also showed that these gaze-cueing effects did not occur for subjects involved in memorizing face images, "a task that is entirely irrelevant to the direction of gaze of observed faces" [85, p. 96]. Thus, whether the gaze of other faces will automatically draw attention to them (or to the location they are gazing at) depends on the kind of task someone is already involved in. The understanding of the precise conditions when such an automatic shift can be expected is subject to future research.

Gaze in Social Interactions

Argyle and Dean [10] proposed the *intimacy equilibrium model* as a framework to understand the relations between different aspects of (nonverbal) behaviors in social interactions. It states that in an interaction between two people, the amount of eye contact is influenced by both approach[11] and avoidance[12] forces. Together with other aspects of interactive behavior (physical proximity, intimicy of topic, amount of smiling, etc.), it contributes to a certain *intimacy equilibrium* felt by the two persons, where each one tries to maintain a level of intimacy where she or he feels comfortable [10]. For instance, if one person moves closer, the other one might reduce the amount of eye contact to compensate for that in order to avoid an undesired high intimacy level. Most of the research discussed below supports such a model or is at least compatible with it, nevertheless there are also studies questioning it. Kendon's [276] view of emotionality regulation by amount of mutual gaze in social encounters is largely compatible with the intimacy equilibrium model. Exline *et al.* [147] and also Schulz and Barefoot [447] found and inverse relation between amount of eye gaze and intimacy of conversation topic. Evidence for an increased amount of mutual gaze at greater distances of interactors was presented by Argyle and Dean [10] and confirmed by Argyle and Ingham [12] and others (e.g. [396, 95]); Kendon found an inverse relation also between similing and amount of gaze [276].[13] However, several other studies (e.g. [73, 3, 397]) yielded different results which indicates that the relation between distance and gaze is complex and depends

[10] However, when cues (e.g. arrowheads) are presented at the periphery instead of the visual center, they can automatically shift the attention to the cued location [262]. Also, if the time between cue appearance and target presentation is long enough, conscious attention shifts to the cued location can occur [262, 405].

[11] e.g. need for feedback or affiliation

[12] e.g. fear of revealing inner states or rejecting responses

[13] However, this result is based on observations of an interaction of one female-male pair of subjects only.

on various other factors. Furthermore, Exline [149] and also Breed [41] showed that instead of compensating an increased amount of gaze or intimacy by interlocutors, some people tend to increase their gazing resp. make their behavior more intimate as well. Argyle also admitted that response matching in terms of gaze and other aspects of interactive behavior have been found [8, 9]. According to Chapman [73], one problem of the intimacy equilibrium model is the implicit assumption that the level of intimacy remains constant during an interaction.

Kendon [276] investigated in detail the gazing behavior during speech. During a conversation, the interlocutors adjust to each other with respect to the amount of eye contact. In general, people look more while listening than while talking (a result that has been confirmend by others as well [376, 148, 128]), also the gaze patterns differ between listening and talking. A common pattern is to look away around the beginning of a long utteracnce, and to look at the interlocutor again at its end. Kendon [276] suggested this might reflect the need for concentration at the beginning and a request of response or attempt to offer the floor at the end of a long statement. He also interpreted the averted gaze he observed during slow or hesitant speech as aid for concentration while uttering complex phrases. Furthermore, both eye contact and averted gaze seem to serve several monitoring and regulatory functions, depending on the exact moment of occurrence and interaction context.

Several studies investigated the influence of the interaction situation on the gazing behavior. Argyle and Graham [11] investigated subjects who planned holidays together and found that the presence of a map—a task-relevant object—shifted the major amount of gaze from the interactants to the map, even more when it was a complex map. The participants of Exline and Winters' [152] experiments gazed less when they performed a difficult task (compared to performing an easier task). According to Kendon [276], a possible explanation is that the reduced visual input due to less gazing aids the concentration on the task. By contrast, a study of Dicks [114] (according to Argyle and Cook [9]) showed that the amount of person-directed gaze of friends playing a board game was no less than in normal conversation, despite the game requiring attention to the board (nevertheless the gazing patterns differed). Exline [148] investigated the gazing behavior of small groups of three people, where the amount of gaze and especially mutual gaze was lower then in typical dyadic interactions. He also found that compared to a collaborative task, a competitive task seems to inhibit mutual glances for people with high affiliative needs and enhance them for low affiliators.

The gazing behavior depends on interpersonal attitudes in many further aspects. In the experiments of Efran [131], people looked more at interlocutors that gave affirmative responses (smiles) and at people of higher status (senior students vs. freshman). There is also evidence for the intuitively convincing assumption that people look more at other people they like [9, 153], although Argyle and Cook [9] pointed out that there is almost no data from people with well-established relationships available, nearly all data comes from short experimental encouters. These authors [9] also discuss several studies examining the partly complex dependencies between gazing behavior and aggression, dominance, shame, embarrassment, sorrow and deception (e.g. [358, 362, 151]). Furthermore, several other emotions were investigated. For instance, people were found to look less when talking about sad incidents [150]. Insulted, angry people react very individually, some look less, others more at the insulting person [376]. Argyle and Cook [9] summarized several studies concerning extraversion and gazing behavior and concluded that their findings are rather conflicting, albeit every study found some relation. Generally, the gazing behavior varies a lot between different people (e.g. [236, 277, 324, 276, 376]), but shows a certain stability for an individual person [277, 324].

Thus, both the interaction situation and the personalities of the involved people significantly influence the gazing behavior, though their precise contribution is a largely open question [9]. Women generally appear to look more at each other than men do [148], also various other differences in gazing patterns between women and men were found in many studies [148, 10, 12, 73, 3]. There are also several intercultural differences, for instance regarding the relation between distance and eye contact [10] and the amount of eye contact in general [514]. Which gazing behavior is appropriate in which situation is largely determined by a set of culture-specific social rules, according to the discussion of Nielsen [376] and Argyle and Cook [9]. An example is the "civil inattention"[14] towards strangers in public places [189].

A few works investigated eye gaze in human-robot interaction. Vollmer *et al.* [506] and also Lohan *et al.* [332] compared the gazing behavior in adult-adult, adult-child and adult-robot interaction. They found a lower amount of eye gaze to the robot than to human interaction partners and suggested the lack of feedback by the robot as likeliest cause, for which they presented some evidence [332]. Hinte and Lohse [219] investigated the reasons why subjects shifted their gaze away from the robot to the experimenter and also found a lack of response as most common cause, among several others.

Conclusion

The gazing behavior of people in social interactions is very complex and depends heavily on interaction context (task to perform, topic of conversation, etc.), personalities of the involved people, interpersonal attitudes, and also culturally learned social rules. Many details regarding the interrelations of these influencing factors are yet unknown. In their research review, Langton *et al.* [312, p. 57] still list *"How does context influence the perception and interpretation of gaze?"* as an outstanding question for future research; other researchers (e.g. [276]) pointed out the necessity of much future research as well.

The research literature discussed above suggests a wide range of functions eye gaze can serve in social encouters, always depending on the particular situation. During speech, gazing at the interlocutor might signal attention or agreement [376], indicate a request of feedback [276], regulate turn-taking [236], or express certainty [376]. Looking away might indicate concentration [276] or dissatisfaction [376], accompany uncertainty [376], or constitute an attempt to hold the floor [276] or to conceal emotional reponses [376]. Kendon [276] further distinguishes *monitoring*[15] from *regulatory*[16] and *expressive*[17] functions and hypothesizes that eye contact can have two opposed purposes: affiliate with somebody and also challenge somebody. The first purpose is evident from studies concerning interpersonal attitudes and the intimacy equilibrium model [10]. Examples for the second purpose are looking at someone to warn her resp. him (of the consequences of inappropriate behavior) [189] and extended eye contact accompanying a "battle for floor" during a conversation [276]. However, prolonged eye contact can also have other meanings, for instance signalling non-understanding [369]. Further functions of eye gaze in social interactions include looking for approval [131], exerting social control [288], signalling willingness to start a social encouter [189], facilitating learning and tutoring (e.g. [176, 389]), and many others [288]. Also various mental states have been attributed to the appearance of the eyes, most notably being thinking when looking upward [20]. For a

[14] One usually does not look directly at strangers, especially in narrow places such as elevators.
[15] observe the behavior of the interaction partner to get information
[16] regulate the interactive behavior, e.g. turn-taking
[17] expresse ones own feelings or attitudes

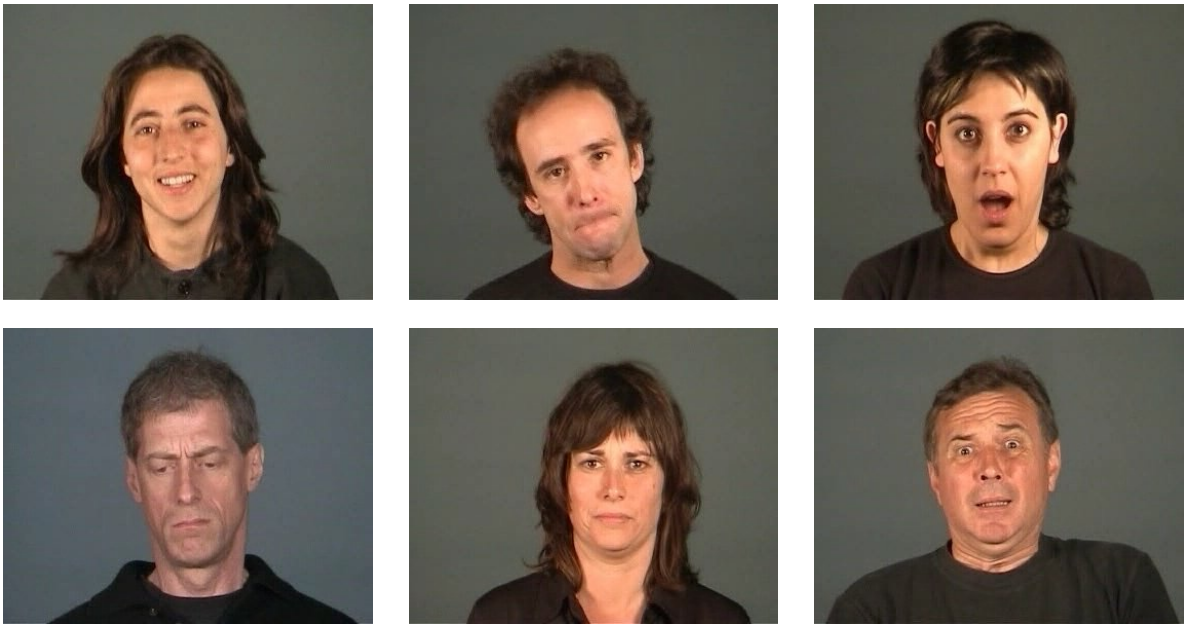


Figure 2.3.: Example displays of facial expressions of *basic emotions* according to Ekman [133]. Top row: happy, angry, surprised. Bottom row: sad, disgust, fear. Please refer to Sec. 2.2.3. The images are taken from the DaFEx database [24].

deeper discussion of possible functions and the complex role of eye gaze in social interactions in general, please refer to the book of Argyle and Cook [9] and the research review of Kleinke [288].

2.2.3. Facial Expressions

Facial expressions received a great amount of research attention in recent decades. The *Facial Action Coding System* (FACS) developed by Ekman and Friesen [139] is the most widley used technique to encode and represent facial expressions. A facial expression is decomposed into a set of *Action Units* (AUs) that are directly related to facial muscle movements, thus a representation in terms of AUs describes the visual appearance of a facial expression and does not attribute a specific meaning to it. Nevertheless it can be used as a solid basis for a subsequent interpretation.

In social interactions, the interpretation of a facial expression is more important than its visual appearance, therefore we focus on this interpretation in the following sections. We discuss facial expressions as emotional displays and an alternative view that emphasizes the communicative meaning in the following two sections. The last section focuses on spontaneous facial displays and contrasts them with posed facial expressions.

Emotional Facial Expressions

One basic question is whether emotions and associated facial expressions are universal and innate, or culture-specific and learned. A widely recognized answer is given by Ekman's

neuro-cultural theory of facial expressions of emotion [133]. It states that some emotions are universally tied to distinctive movements of facial muscles by a *facial affect program* and thus produce particular facial expressions. Nevertheless there are cultural variations regarding the elicitors of emotions, social display rules for facial expressions, and consequences of emotional arousal, all of these are socially learned. Ekman [133] reported several studies that provide some evidence for this view, particularly showing that fear, anger, sadness, disgust, surprise, and happiness can be recognized by people of five different literate cultures from the same set of images; and also to some significant degree by people of two preliterate cultures. Fig. 2.3 shows posed example displays of these emotional facial expressions.[18] In later work [134], he admitted that the evidence of surprise is not as firm as for other emotions, and suggested awe, contempt, embarrassment, excitement, guilt, interest, and shame as additional candidates for basic emotions, leaving that question for future research. Furthermore, he discussed characteristics of basic emotions that might distinguish them from other affective states or non-basic emotions. Izard [247] largely agreed with Ekman's view, but named the slightly different set of interest, joy, surprise, distress, anger, disgust, contempt, fear, shame, and guilt as *fundamental emotions*. In addition he proposed the existence of fairly stable *affective patterns* of emotions where several fundamental emotions interact to form further affects, for instance anxiety, depression, love, and hostility. Other researches suggested in part different sets of discrete emotions (e.g. [13, 402]). More recently, Tracy and Matsumoto [489] presented evidence for innate facial expression of pride and shame as reactions to success and failure. Matsumoto and Kupperbusch [343] found expressional differences depending on personality.

Russell [429] questioned the evidence for universial recognition of emotions from facial expressions gained from the studies of Ekman *et al.* [143, 133, 141], Izard [246] and others [377, 38, 125, 349]. He criticized the experiments on several levels and discussed potential problems that might have distorted the results (e.g. forced choice experiments, subject selection, subjects possibly knowing about the universality thesis, within-subject design, previewing and order of image presentation, preselection of images, and posed expressions) and suggested several alternative interpretations (e.g. bipolar dimensions, response to a situation, different facial expression categories, subjects might rather "solve a puzzle" than truly interpret facial expressions). Ekman [136] rejected this criticism and argued that most of the potential problems did not actually occur as great care was taken to avoid them, other addressed issues were not regarded as beeing problematic. Similarly, Izard [248] defended his position of innante and universal facial expressions. Nevertheless he admitted that Russell [428] showed "that we probably have not yet determined the exact number or best names of emotions with universal expressions." [248, p. 297] However, these objections satisfied Russell very partially only. He was especially not convinced of the superiority of Ekman's interpretations over several alternatives [430].

One widely used alternative are *dimensional models* that regard emotions as varying along bipolar, nearly independent dimensions. Mehrabian and Russell [359] presented pleasure, arousal, and dominance as the three fundamental emotional dimensions, largely based on *semantic differential* studies [384, 385, 465, 56]. In later work, Russell [425] reviewed several studies and presented evidence for pleasure and arousal dimensions, whereas the evidence for dominance was not as clear; Russell considered it no longer an affect dimension later on [426] (although further, nonaffective dimensions apparently exist). He also defended this model against methodical criticism (e.g. [197]) and studies suggesting monopolar dimensions (e.g. [484, 357]) by criticizing the response format and correcting for a thereby introduced

[18] These images are not part of the image set Ekman [133] used in his experiments.

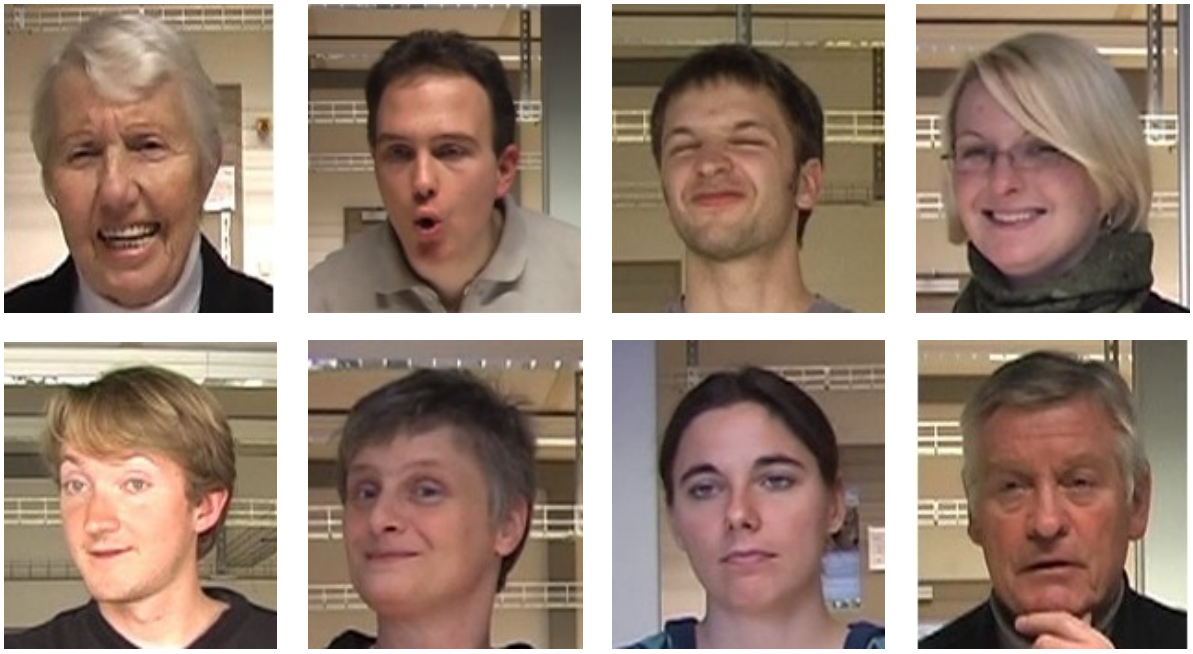


Figure 2.4.: Example displays of *communicative* facial expressions that occured spontaneously in real human-robot interactions. Please refer to Sec. 2.2.3.

bias [426], though there is some evidence favouring monopolar interpretations, especially showing that positive and negative affect can occur simultaneously (e.g. [517, 356, 454, 237]). Further research led to the *circumplex model of affect* [427, 404], where the pleasure and arousal dimensions are viewed as systematically interrelated rather than independent, such that a large number of affects roughly resemble a circle along these two dimensions. Russell *et al.* [432] further presented some evidence showing that this model is appropriate for a broad range of cultures. Bradley and Lang [39] developed the *self-assessment manikin* as a simpler alternative to the semantic differential to access affective responses and also proposed a pleasure-arousal-dominance model.

Besides the models discussed above, there are several other views on emotions. The *appraisal theory* regards someons's evaluation resp. appraisal of a situation as determinant for the kind of emotional reponse that is felt [443]. Other researchers linked components of facial expressions to emotion dimensions [464] (according to [215]) or discrete emotions [63] or proposed a determining connection between emotions and the readiness to take specific actions [175]. Izard [249] recently pointed out that there is still no commonly accepted definition of "emotion", despite broad agreement on several aspects. He suggested that researchers should contextualize their understanding of emotion to clarify its meaning. Widen and Russell [519] added on that by emphasizing the difference between an every day and a scientific concept of emotion.

Communicative Facial Expressions

Fridlund [169] presented the *behavorial ecology view* of faces which is very different from the emotions view discussed above. Facial expressions are regarded as communicative signals

that enhance social interaction rather than external displays of internal emotions. Fig. 2.4 depicts several examples of such communicative facial displays. However, no *prototypical* facial expressions are proposed, as the meaning of a facial display depends heavily on the context. This view is supported by human audience effect studies. Generally, smiles most often occur in social contexts [14, 407]. Kraut and Johnston [295] reported that bowlers' smiles were much more related to social interaction with the people around than to scoring a strike or spare. Similar results were obtained for people experiencing good or bad weather and, to a weaker degreee, for fans watching a hockey game. Even olympic gold medal winners smiled frequently only during the interactive parts of the awards ceremony, in spite of their apparent intense happiness throughout the whole ceremony [161]. Fridlund *et al.* [170] reported that people smiled more when imagining high-sociality situations compared to low-sociality ones and that the degree of smiling was little related to their happiness. Fridlund [168] also showed that people watching an amusing video smiled more when a friend was present, or even when they were told that a friend was in a room nearby, each compared to watching alone. Very similar results were obtained by Chovil [80] for the facial display of subjects hearing about close call events. Bavelas *et al.* [25] presented evidence that the motor mimicry of subjects observing apparently painfully injured victims can reasonly be interpreted as communicative act. Brightman *et al.* [43, 44] found that observing judges could easily tell whether videotaped subjects were eating sweet or salty sandwiches when the subjects were in company, but not when they were alone. Grammer *et al.* [196] presented evidence for social functions (signalling "yes", emphasis of other cues) of eye-brow flashes, for which they found certain movement patterns in three different cultures.

However, there is also evidence that social context can inhibit the display of negative facial expressions (e.g. [287, 294, 250]). Jakobs *et al.* [250] interpreted their results for sad faces as being largely compatible with the idea of display rules as suggested by Ekman [133] and less supportive for the behavorial ecology view [169], but also admitted that the experimental setting might have influenced the subjects against behaving as expected by this view. Ekman *et al.* [138, 134] emphasized that facial expressions also occur when people are alone and not imagining others, which questions a sole communicative role. Furthermore, Ekman [137] defended the emotional view of facial expressions and argued that they were not deliberately made to communicate, although emotions play a role in interaction. Nevertheless other kinds of facial expressions appear to be very important, as Chovil [79] (according to Heylen [215]) observed that hardly 20% of the facial expressions during face-to-face conversation are affective in nature. Chovil and Bavelas [79, 26] also identified several discourse functions of facial displays of both speakers and listeners (e.g. communicate personal reactions, thinking/remembering, "facial shrugs", backchanneling, express sincere appreciation by motor mimicry, signal understanding) and stressed that facial displays need to be interpreted in their verbal and conversational context.

Parkinson [393] compared Ekman's [133] and Fridlund's [169] approaches and reviewed both in terms of theory, evidence, and consequences. He concluded that neither approach can account for all the available evidence: many results cannot be explained by a pure emotions view, the behavorial ecology view covers a wider range of phenomena, but is too imprecise regarding the exact relation of the facial display to social motives and audience effects and cannot explain all emotional displays. Thus, further research should aim at a comprehensive theory of facial movements and state these relations more precisely [393]. Fernández-Dols and Ruiz-Belda saw "an urgent need to avoid hurry" [162, p. 270] concerning theory construction and suggested to systematically collect a large amount of new data before theoretical development should

proceed, because a "commitment to a premature theoretical framework" [162, p. 270] can and actually did delay research progress.

Spontaneous Facial Expressions

It is important to note that spontaneous facial expressions are quite different from posed ones. This is why Russell and Fernández-Dols [431] worried about the large amount of research on posed facial expressions. The goal of the posing person most likely is to make the display easily understandable by the observer [429], which does not necessarily hold for authentic expressions. Then again, naturally occuring facial displays in conversations might be "posed" in some sense and to some degree in order to be expedient as social signals with communicative meaning. Thus, Russell and Fernández-Dols [431] surmised that both posed and spontaneous facial expressions might be two (rarly occuring) extreme endpoints on a continuum.

Ekman *et al.* [142] reported differences in the visual appearance between posed and spontaneous smiles. Fernández-Dols and Ruiz-Belda [162] concluded that the relation between happiness and spontaneous smiles is unexpectedly complex; happiness is neither sufficient nor necessary for smiling. In a series of experiments, Reisenzein *et al.* [415] demonstrated a strong dissociation between spontaneous surprise and facial displays on several levels. Different emotional facial expressions can differ considerably concerning their timing [46]. Ekman [136] suggested that spontaneous emotions should be investigated with videos instead of single images, to capture the temporal dynamics. He also described a certain dissociation between emotion and facial expressions [135].

Several studies suggest that spontaneous facial expressions are *much* more difficult to interprete by human observers than posed ones. Motley and Camden [365] found a huge decrease of recognition accuracy when human subjects judged authentic facial expressions, compared to posed expressions of the same basic emotions shown by the same people. A similar low accuracy for spontaneous expressions was obtained in the studies of Wagner *et al.* [508]. Gilbert *et al.* [188] observed a large discrepancy between the interpretation accuracies of posed and spontaneous reactions to odors. According to Bruce *et al.* [46], at least positive and negative spontaneous facial displays of emotion can be distinguished.

2.3. Conclusion

The extensive discussion of FCSs above shows that display, meaning, and interpretation of head gestures, eye gaze, and facial expressions depend heavily on the interaction context and are very complex and multifaceted, even more so when one considers the interrelations between these signals (e.g. [438, 32, 82, 216, 2, 309, 311, 46]). We draw several conclusions from this discussion which also motivate the work presented in the following chapters:

- In natural, real world interaction situations, almost always a (complex) combination of head gestures, eye gaze, and facial expressions occurs. Therefore, we suggest to consider these three kinds of FCSs not as single modules, but in combination altogether.
- We need to be aware of the major importance of the interaction context when we study FCSs within in a certain scenario, as the concrete kind and setup of interaction is very likely to have an non-negligible influence on the way the involved people display FCSs

(please see Sec. 3.7). Thus, an investigation of FCSs "in general" in a single scenario is hardly possible.

- Due to the complexity and context-dependence, we doubt that a comprehensive, general purpose interpretation of FCSs by robots interacting with humans will be feasible in the near future. Therefore, we think that a pragmatic simplification and focusing on different, specific interaction scenarios will remain necessary and beneficial for the midterm development of this aspect of human-robot interaction.[19] (Thus, the overall interpretation capabilities of a robot rather might arise from the combination of several subsystems that are dedicated to specific tasks and contexts than from one general purpose system for FCS interpretation.)
- To discover which FCSs actually occur in a specific interaction scenario, we prefer a data-driven approach to an a priori modelling, as it might be very difficult to anticipate which FCSs are the most prominent ones in a certain context.[20]
- Due to the complexity, context-dependence, and ambiguity of FCSs, the acquisition of reliable ground truth data for an automatic classification is an important, nontrivial issue. For instance, the judgements of human observers are likely to be subjective and might exhibit a rather low inter-rater agreement (please see Sec. 3.1). We suggest a definition of ground truth in terms of the objectively ascertainable interaction situation instead of the visual appearance of the face, because this circumvents some typical problems. Please refer to Sec. 3.3.
- We suggest an FCS interpretation in terms of broader clusters (e.g. positive vs. negative) instead of finegrained categories, because we expect the former to be applicable in a wider range of situations resp. contexts and thus generalize better to other interaction scenarios. The actual facial displays subsumed by one of the broader categories might still vary between different scenarios resp. contexts, but the category as an interpretation concept would be suitable nevertheless. By contrast, finegrained categories might be specific for a much smaller class of scenarios, not only regarding the appearance of the displayed FCSs, but also regarding the category label as interpretation concept in itself (please refer to Sec. 3.7).

The next chapter presents our approach to FCS recogniton in a specific scenario where these conclusions are considered.

[19] For some specific scenarios, there already exist very good solutions for an automatic FCS recognition. A prominent example is the smile detection feature of most recent digital cameras, which can reliably recognize posed smiles of people facing the camera. However, the recognition of smiles in general, particularly genuine smiles, in unconstrained environments and interaction situations is a yet unsolved problem.

[20] Nevertheless such an a priori modelling should be possible when one can rely on enough previous experiences in the same setting.

3. Valence Recognition

Everything should be made as simple as possible, but not simpler.
— Albert Einstein

The extensive discussion of the research literature about facial communicative signals (FCSs) in the previous chapter shows the necessity of a suitable simplification in our investigation of FCSs in human-robot interaction. We decided for a valence-based approach to FCS interpretation in an object-teaching scenario, which is explained in detail in this chapter. The following Sec. 3.1 briefly addresses our motivation for the choice of the object-teaching scenario, which is described in the subsequent Sec. 3.2. Section 3.3 introduces the valence recognition task and explains the definition of the ground truth data. An evaluation of the facial displays shown by the participants of the study is presented in Sec. 3.4. The human performance in the valence recognition task is investigated in Sec. 3.5, afterwards our investigations are compared to related studies in Sec. 3.6. Finally, Sec.3.7 discusses some aspects of the interaction context.

3.1. Selection of a Suitable Human-Robot Interaction Scenario

In order to find relevant FCSs that actually do occur in typical human-robot interactions, we evaluated videos of two previously conducted user studies. In the first study, several participants showed around a robot in an apartment [334], while the names of several objects were taught to a robot in the second study [335]. Fig. 3.1 shows example images of the recorded videos. Neither study was related to FCSs originally, but both situations constitute realistic human-robot interaction scenarios for a relevant FCS analysis.

Several human raters judged the videos of these studies with respect to FCSs. Not surprisingly, it turned out that typical facial expressions of basic emotions [133] rarely occured. This could by expected according to the discussion of communicative facial expressions in Sec. 2.2.3; it is also in accordance with the experiences of Caridakis *et al.* [62] who investigated automatic facial expression recognition in the context of human-machine interaction. In many cases, the raters found the observed facial displays not very clearly visible but rather subtle and difficult to interpret in terms of exact categories. (More than 50 categories were named by these observers.) The agreement about the best suited category for a particular facial display was often poor, also an expedient and comprehensive set of those categories was not defined.

Nevertheless, the object-teaching scenario was found to be well-suited for FCS studies in general because of the frequent occurrence of FCSs, despite their difficult interpretation in terms of precise categories. However, a classification into broader clusters (e.g. positive vs. negative) achieved much higher agreement among the human raters, which is one motivation for the valence-based recognition approach presented in Sec. 3.5. The videos of the evaluated


Figure 3.1.: Example images from two existing video databases that were evaluated with respect to the occurring FCSs. Left: A person teaches an object to a robot (not shown) [335]. Right: A participant shows an apartment to the robot Biron [334, 204]. Please refer to Sec. 3.1.

object-teaching study do not contain close up views of the participants' faces, moreover the camera perspective provides no frontal view on the faces. Since frontal face images in a sufficiently high resolution are required or at least desirable for later automatic analysis, it was necessary to conduct a new object-teaching study, which is described in the next section.

3.2. Object-Teaching Scenario and User Study

We conducted the object-teaching user study with 11 subjects (five female and six male) ranging from 22 to 77 years in age, nine of them had never interacted with the robot before. The participants were instructed to show several manipulable objects to the robot "Biron"[1][204] and to teach the objects' names. Furthermore, they should validate that the robot had actually learned the objects. It was not specified how the objects should be termed and presented (e.g. pointing to them, taking them in hand, etc.), but the subjects were asked to interact with the robot at their convenience, in order to perform the given task. They were aware that the robot understood speech and could see them. (Please refer to App. A.1 for a copy of the instructions the subjects received.) The robot interacted with the subjects by voice production and movements of its pan tilt camera (most of the time focusing on either an object or the subject's face).

We performed a Wizard of Oz study where Biron was remote controlled to determine exactly its behavior (when to recognize the object correctly, when to misunderstand the subject, what to say, and where to look). Of course the subjects did not know this, but assumed autonomous operation of the robot whose object recognition capabilities were to be evaluated, whereas in fact the study was about provoking authentic, spontaneous FCSs for later analysis. The subjects were not aware that FCSs were of any interest during the study, which we regard as a necessary prerequisite when authentic FCSs are investigated. A pre-study using the same setup confirmed that they were likely to display FCSs spontaneously nevertheless.

[1] BIelefeld Robot CompaniON

Per person, two counterbalanced sessions were performed: a "good" one where Biron termed most of the objects correctly, and a "bad" one where Biron misclassified the majority of objects. A session lasted between 6 and 15 minutes (about 10 minutes on average). For every person, the "bad" session took longer than the "good" one, because the subjects spent a significant amount of time correcting the misclassifications of the robot (about 8 resp. 11 minutes for "good" resp. "bad" sessions on average). Between the two sessions of a subject, the objects were exchanged to make the subjects believe that the robot's recognition performance on another object set was to be evaluated and thus give a seemingly plausible explanation for its performance differences. During each session, videos were recorded from three different perspectives as shown in Fig. 3.2. One stationary camera in front of a table with nine objects recorded the whole scene, showing the robot Biron on the left and the subject on the right. Another stationary camera was placed right behind Biron to record the face of the subject during the whole experiment. Additionally, the videos taken by Biron's pan tilt camera were stored.

3.2.1. Video Database Description

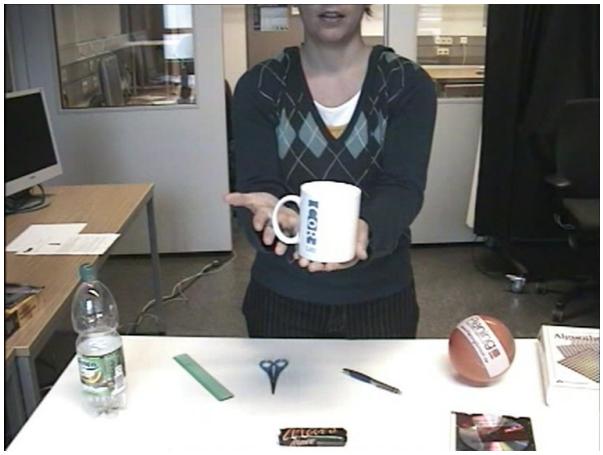
To support the evaluation of the interaction videos in terms of FCSs, all videos recorded by the stationary face camera were manually annotated. An inspection of these videos showed that the object-teaching scenes are highly structured. This motivates the subdivision into four subsequent phases, which were annotated in addition to the transcription of the speech of both subject and robot:

1. **present phase:** The subject presented the object to Biron and said its name or asked for the name.
2. **waiting phase:** The subject waited for the answer of the robot (not mandatory).
3. **answer phase:** The robot answered the subject, for instance, by classifying the object or asking a question.
4. **react phase:** The subject reacted to the answer of the robot.

In many cases, the interaction scenes overlap, because a part of the react phase of one scene might be part of the present phase of the next scene. This usually occured when Biron misclassified an object and the subject reacted by saying the correct object name and presenting the object again, which also started the next interaction scene. The exact times of the phases were sometimes ambiguous (especially the end of react or present phases). To achieve consistency nevertheless, all scenes were annotated according to a predefined coding scheme. Each object-teaching interaction scene was classified into one of the following categories, depending on the answer of Biron (example answers in parenthesis[2]):

- **success:** Biron said the correct object name. *("So, this is a book." after the subject taught the object name or "This is a book." after the subject asked for the object's name)*
- **failure:** Biron said an incorrect object name. *(same answer structure as in the success case)*
- **problem:** There was a communication problem, but Biron did not say any object name. *("I don't know the object.", "I don't know the word.", "I don't know.")*

[2]These answers were translated into English. Please refer to App. A.2 for a comprehensive list of the German utterances Biron used during the study.



(a) Biron's pan tilt camera

(b) Stationary face camera

(c) Stationary scene camera

Figure 3.2.: The object-teaching video database contains videos showing human-robot interactions from three perspectives [305]: (a) the videos recorded by the robot's pan tilt camera, (b) the subject's face recorded by a stationary face camera that was placed right behind the robot, and (c) the whole interaction scene recorded by a stationary scene camera. Please refer to Sec. 3.2.

scene type ↓ | subject →	01	02	03	04	05	06	07	08	09	10	11	total
success	15	17	32	20	16	15	25	32	13	12	24	221
failure	18	11	21	17	16	13	31	26	24	12	35	226
problem	12	6	14	4	4	2	4	23	5	0	2	76
vague	6	1	1	2	1	0	6	5	0	1	1	24
clarification	26	16	16	16	14	10	23	22	19	12	21	195
abort	0	0	0	0	0	3	3	1	0	1	1	9

Table 3.1.: Number of object-teaching scenes of six different categories for the 11 subjects in the video database. Please refer to Sec. 3.2.1.

- **vague:** Biron claimed to understand, but did not say any object name. *("I have seen the object.", "This is interesting.", "I like it.")*
- **clarification:** Biron asked a clarification question. *("Pardon?", "I could merely understand you partially. Can you repeat this, please?", "Did you show me the object before?")*
- **abort:** Biron did not answer in a reasonable period of time, thus the subject aborted this interaction and taught a new object.

There were only very few cases where an interaction scene did not match any of these categories. Those scenes were omitted. In addition to the interaction phases, the period of time the robot said an object name (in *success* and *failure* scenes) was annotated. This information was used in the user study that investigated the human valence recognition performance as reported in Sec. 3.5. In total, 751 interaction scenes were annotated. From Tab. 3.1 it can be seen that the study focused on *success* and *failure* scenes. Moreover, *clarification* scenes were frequent, whereas the remaining scene categories occured far less often. This focus was chosen in view of the targeted evaluation in terms of FCSs. In a pre-study, we tested also additional behaviors of Biron which resulted in more interaction scenes of other categories. However, it turned out that a strong focus on *success* and *failure* scenes was necessary to yield a sufficient number of these scenes for FCS evaluation.

3.2.2. Comments from the Participants

After the interaction with the robot, we asked the participants to fill out a questionnaire about the experiment.[3] Ten of the eleven subjects attested Biron that it was "good" at learning the objects, whereby seven subjects explicitly mentioned differences between the two trials. Five people commented on the slow interaction speed of Biron, especially for its feedback, leading to an "unnatural" or "stiff" feeling of the interaction; one person wrote that one gets used to it nevertheless. Also five subjects got the impression of an autonomously acting robot, whereas Biron's limited vocabulary and relatively restrictive interaction schemes were addressed by five other participants, which obviously hindered getting an autonomous impression of Biron. (However, they compared Biron to humans, not to other robots.) Three persons also mentioned Biron's problems in understanding their speech (which in fact was on purpose due the study design). No one hinted the suspicion of a remote-controlled robot

[3]Please see App. A.3 for a copy of this questionnaire.

resp. Wizard of Oz study on the questionnaire, this and the "true" purpose of the study with respect to FCSs was told at the very end after the completion of the questionnaire. Four subjects stated they had fun during the experiment,[4] three found it interesting, two good, two pleasant, one cool, and one strange. One person wrote one gets the impression of playing with a little child.

3.2.3. Interactive Behavior

As part of her dissertation, Manja Lohse [333, pp. 95–125] analyzed various aspects of the interactive behavior of the participants. To avoid sequence effects, the two sessions each subject performed were counterbalanced (please see Sec. 3.2). However, Lohse did not find any sequence effects in her analysis, so it seems not matter whether the subjects performed the positive or the negative session first. The interaction structure with respect to the duration of present, wait, answer, and react phase was very similar in both cases. Also the utilized gestures did not differ significantly between positive and negative sessions, they rather appeared to be a matter of personality. Lohse pointed out that the robot gave no feedback on gesture level (as it did not use gestures at all), which might be a likely cause why the subjects did not adapt their gesturing behavior. She also investigated the gazing behavior of the participants, which we discuss in Sec. 3.4.2. Furthermore, she statistically proved the intuitively expected finding that positive and negative sessions can be differentiated based on the speech of the subjects. Please refer to her dissertation [333, pp. 95–125] for the full details of her investigations.

3.2.4. Deployed Software

In order to perform the remote control of the robot, we developed a Java application that allowed to move the pan tilt camera and to let Biron utter several predefined and custom phrases via a simple graphical user interface (GUI). This software uses the *BonSAI*[5] library [458] as back-end to access the robot's camera and speech production system. The videos of Biron's pan tilt camera were recorded through the image grabbing system of the *iceWing*[6] [337, 336] framework. (The scene camera and the stationary face camera were ordinary cameras recording on MiniDV cassettes.) The transcription of speech and annotation of scenes and interaction phases was done with *ELAN*[7] [346, 521].

3.3. Valence Recognition Task and Ground Truth Data

This section explains how the ground truth data is defined for the valence recognition task in our object-teaching scenario that is investigated in this thesis. The usual practice in visual recognition resp. classification tasks is to define the ground truth labels in terms of the *visual appearance* of the objects under investigation. This is perfectly fine for typical object recognition problems, for instance, as an accurate label can usually be easily assigned by a human who can determine the correct class of an object without difficulties. However, in

[4] However, one of these persons doubted that Biron actually had fun in the experiment, despite saying so in its farewell phrase.
[5] Bielefeld Sensor and Actuator Interface
[6] A graphical plugin shell optimized for image processing
[7] A tool to create complex annotations of audio and video resources

case of FCSs, the situation is different. As discussed in Sec. 2.2, FCSs and the way people use them in interactions are very complex and multifaceted, there are different, competing views about their nature an the best way to consider them. Concretely, this means that for given displays of FCSs, very often not only the class it should be assigned to is difficult to determine, but also the set of classes that should be used as interpretation categories at all. The second problem is usually solved by either using a (sub-)set of established categories (for example *basic emotions* [133] in case of facial expressions)—thus deciding for a particular psychological model—or by pragmatically defining categories that seem to match the concrete data and context at hand best. To solve the first problem, the assignment of class labels to data instances, typically one of the following approaches is taken:

- The subjects are asked to display FCSs of given categories on request. This has been done in many studies, a prominent example is the DaFEx database [24]. While this method has appealing advantages, most notably a comparatively safe and well-defined acquisition of labeled data, and has already been very useful in computer vision research, it also has serious drawbacks: The FCSs are not *authentic* and *spontaneous*, but *posed*. This is most likely an issue, because the differences are usually very prominent (please see Sec. 2.2.3). This can be moderated if the subjects are professional actors who are trained to pose FCSs "naturally", as it is the case for the DaFEx database, but the differences are unlikely to vanish completely. Furthermore, the display of FCSs depends heavily on the interaction context, which is very difficult to consider appropriately when these signals are posed.

- Human raters judge video recordings of interactions where the subjects show authentic, spontaneous FCSs. This avoids the problem of posed FCSs, but relies on the necessarily subjective and often ambiguous impression of the raters. The reliability can be increased by having several raters judging the videos in parallel and accepting only those FCS displays where a majority agrees on. However, this causes a very high expenditure of human labor and could lead to a significant amount of rejected data instances with too poor agreement.[8]

- The subjects are interviewed about the intended meanings of their facial displays. This is less common than the first two approaches, probably because of the practical problems: When the interview takes place after the experiment is over, it might be very difficult for the subjects to remember the intended meanings of several facial displays in particular situations. On the other hand, interrupting immediately after every interesting situation is likely to disturb the experiment or influence the subjects in an undesired way (e.g.[325]).

To cope with these problems, we used a different approach where the ground truth data is defined in terms of the objectively ascertainable *interaction situation* instead of the visual appearance of the face. In our object-teaching scenario, we focused on *success* and *failure* scenes, which are defined by the (either correct or wrong) answer of the robot when it classified an object. The FCSs displayed in these situations are treated as examples for the respective class: *success* or *failure*. In a sense, this is an inverse approach: instead of trying to find the correct ground truth labels for given facial displays, we look for FCSs in a given situation with implicitly given ground truth data.

[8] As written in Sec. 3.1, we experienced a rather poor agreement in a pre-study of this work.

While this approach yields reliable ground truth labels, it faces another problem: As the definition of these labels is solely based on the (outcome of the) interaction situation and independent of the visual appearance of the face, their is no guarantee that a meaningful FCS is displayed at all. However, the studies of Barkhuysen *et al.* [18] and also our pre-study suggested that usually a meaningful display occurs. The evaluation in the following Sec. 3.4 shows that this actually is the case for this object-teaching study.

Thus, the research question investigated in this thesis is not the standalone interpretation of FCS in itself (as in most research on facial expression recognition), but their interpretation as *feedback* about the interaction in terms of *valence*, and the question to which degree this feedback can be gained from FCSs at all. One can regard this as interpretation on *pragmatic* level, while the former is on *semantic* level. For all the following investigations, we did not use the complete scenes (present, wait, answer, and react phases), but extracted a subpart of the associated videos from the stationary face camera (please see Fig. 3.2), starting near the end of the answer phase, exactly when Biron started to say the object name, and ending at the end of the react phase. This starting point of the videos was chosen because it is the first moment from which the subject could know whether the answer of the robot was correct or not. Hence, this part of the interaction scene appears to be the relevant one for FCS analysis regarding the feedback as discussed above.

3.4. Display of Facial Communicative Signals

We visually inspected all *success* and *failure* videos of all subjects to get an impression about the kinds of FCSs they displayed during the interaction with Biron. Please refer to Fig. 3.3 for typical example displays of these FCSs, and to Tab. 3.2 for statistics about their occurrence. The next three sections discuss that for head gestures, eye gaze, and facial expressions individually, partially with the targeted automatic recognition in mind.

3.4.1. Head Gestures

As intuitively expected, the majority of head gestures were head nods and shakes. In 58.0% of the *success* scenes, the subjects showed a head nod, while a head shake was displayed in 29.1% of the *failure* scenes only. In both cases, the variance between different subjects was very high (Tab. 3.2, rows 1 and 2). For instance, subject 09 showed one (weak) nod only and no shakes at all, while subject 03 used head nods and shakes in more than three-fourths of all interactions. A head shake was never shown in a *success* scene, and a head nod occurred only once in a *failure* scene: at the last of four misclassifications in a row, subject 07 told Biron that they need to practice more and nodded to confirm this statement.

Thus, head nods and shakes appear to be reliable indicators for *success* and *failure*, in principle. However, 69.0% of the nods and 56.0% of the shakes were very weak, meaning that either the upward or downward movement (or both) in case of nods resp. the sideward movement for shakes is very small, usually covering very few pixels in the videos only (Tab. 3.2, rows 3 and 4). Again, the variance between subjects was very high. This might be a problem for an automatic detection of these head gestures, as these small movements are likely to be in the range of typical matching inaccuracies of common computer vision approaches. For head nods, this is aggravated by the fact that the recognition of the upward or downward movements alone

Row	FCS Display ↓ | subjects →	01	02	03	04	05	06	07	08	09	10	11	mean	SD
1	**head nods / success**	60	76	78	65	38	73	64	25	8	83	67	58.0	24.0
2	**head shakes / failure**	28	0	76	53	6	38	84	4	0	8	23	29.1	30.3
3	**weak head nods**	56	85	60	69	17	36	59	87	100	90	100	69.0	26.7
4	**weak head shakes**	80	-	38	22	100	40	12	100	-	100	13	56.0	38.7
5	**upward moves / failure**	28	58	57	29	19	54	42	62	83	92	57	52.8	22.4
6	**weak upward moves / failure**	20	100	42	80	67	29	62	56	35	73	80	58.4	24.8
7	**downward moves / failure**	72	58	52	24	38	54	26	38	67	83	54	51.5	18.8
8	**weak downward moves / failure**	15	100	36	100	50	29	63	30	38	60	68	53.5	28.0
9	**gaze at robot / success**	100	100	100	95	50	100	96	88	92	42	100	87.4	21.1
10	**gaze at robot / failure**	100	100	100	100	94	92	100	100	100	83	100	97.2	5.4
11	**gaze at object / success**	0	0	0	10	13	0	16	25	0	8	0	6.5	8.6
12	**gaze at object / failure**	0	0	38	53	31	46	39	12	17	25	17	25.2	17.9
13	**gaze downwards / success**	100	82	100	90	81	67	60	78	100	92	88	85.2	13.3
14	**gaze downwards / failure**	83	67	48	24	19	38	32	27	88	83	23	48.3	27.0
15	**gaze elsewhere / success**	0	0	0	50	94	20	0	9	33	58	0	24.1	31.5
16	**gaze elsewhere / failure**	0	8	29	76	100	54	6	27	33	83	26	40.3	33.6
17	**talking / success**	100	100	97	100	100	93	100	94	92	100	100	97.9	3.3
18	**talking / failure**	100	100	90	100	100	100	100	100	92	92	100	97.6	4.1
19	**pronounced talking / success**	20	82	0	65	6	13	8	0	0	8	0	18.5	28.3
20	**pronounced talking / failure**	83	92	29	88	69	38	26	73	92	42	94	65.7	27.6
21	**laughter / success**	0	6	0	15	0	80	36	0	0	0	13	13.6	24.7
22	**laughter / failure**	17	33	0	12	6	31	29	0	29	42	49	22.5	16.6
23	**affirmative expression / success**	0	0	69	15	0	13	40	9	0	0	0	13.3	22.1
24	**affirmative expression / failure**	0	0	52	18	0	0	35	8	0	25	3	12.3	17.8
25	**negative expression / failure**	0	25	43	12	31	31	10	50	13	50	6	24.5	17.9
26	**weak negative expression**	-	0	78	50	80	0	0	85	100	33	50	47.6	38.1

Table 3.2.: Amount (percentage) of interaction scenes where specific FCSs occurred, for each subject of the object-teaching study, and mean value and standard deviation (SD) over all subjects. Upper block: head gestures, middle block: eye gaze, lower block: facial expressions. Please refer to Sec. 3.4 for an explanation and discussion of each row.

(which might be easier because very often one part is more pronounced than the other) cannot be expected to be sufficient, as both upward and downward movements occurred frequently also in *failure* scenes, without being part of a nod.[9] In many cases, a person moved the head (slightly) upward to prepare a pronounced uttering of the correct object name (and downward afterwards again in some cases). For head shakes, the situation is somewhat easier in the sense that sideward movements were displayed also outside of shakes, but usually co-occurred with downward movements in the process of looking around at the table with objects and showed a wide movement range, so that the confusion risk might be lesser in this case. Moreover, a lateral head position (looking elsewhere into the background) appeared several times during the answer of Biron, most notably for subject 04. The subjects probably did this in order to focus on Biron's verbal answer (please see the discussion of Kendon's work in Sec. 2.2.2). Other types of head gestures were found in about 2% of the interactions only. These include head movements that do not appear to carry a communicative meaning, for instance fast movements with the apparent purpose of moving hair out of the face.

3.4.2. Eye Gaze

We evaluated whether or not the subjects looked at the robot, the object, down to the object table, or elsewhere during an interaction with Biron. Gazing at the robot was very common, the subjects did this in 87.4% of the *success* and in 97.2% of the *failure* scenes (Tab. 3.2, rows 9 and 10). Though the differences between *success* and *failure* were notable for some subjects (namely 05 and 10), they were not significant on average.[10] Glances to the object that was currently taught occurred significantly less often in *success* than in *failure* scenes (6.5% vs. 26.5%, Tab. 3.2, rows 11 and 12). Gazing downwards to the table with objects was significantly more typical for *success* than for *failure* (85.2% vs. 48.3%, Tab. 3.2, rows 13 and 14). In contrast, the amount of interaction scenes where the participants gazed elsewhere did not vary significantly (24.1% vs. 40.3%, Tab. 3.2, rows 15 and 16). The variance between different people was particularly high in the last case, but also considerably large for the gazing targets discussed before. Hence, gazing at the object is some indication for *failure*, likewise is gazing downwards for *success*. However, although significant, the differences in the respective amounts are not large enough to base a classification solely on gaze (the subjects gazed downwards also in almost half of the *failure* scenes, for instance). Moreover, we observed some individual, specific behaviors. For instance, subject 07 often closed the eyes for about one second in her reaction to *failure*, and subject 09 used very frequent gaze shifts, particularly in *failure* scenes.

In her dissertation, Manja Lohse also evaluated the gazing behavior of the participants of this object-teaching study [333, pp. 117–119]. She found that the mean duration of glances

[9] upward movements in 52.8% of *failure* scenes (58.4% of them weak), downward movements in 51.5% of *failure* scenes (53.5% of them weak), Tab. 3.2, rows 5–8

[10] Throughout this paragraph, the significance of the discussed differences between *success* and *failure* was tested by both a two-tailed t-test and a Wilcoxon rank sum test. These double tests were performed because preceding Shapiro-Wilk tests for normal distribution yielded mixed results, i.e. a normal distribution was rejected for some data (namely rows 9, 10, 11 and 15 in Tab. 3.2) at 0.05 significance level while it was not rejected for other data (namely rows 12, 13, 14 and 16 in Tab. 3.2)). To achieve conclusive results nevertheless, both the t-test (assumes normal distribution, commonly used in the literature) and the Wilcoxon rank sum test (does not assume normal distribution) were performed. For all data discussed in this paragraph, both tests yielded the same basic result ($p < 0.01$ in both tests for all significant results, $p > 0.15$ in both tests for all non-significant results).

at the robot was rather high (compared to typical human-human interactions), substantially influenced by relatively long *wait* phases where the person usually gazes at the robot while awaiting its answer. Such long glances would probably not be socially appropriate in conversations with humans. In positive sessions,[11] there were longer glances at objects, while longer glances at Biron and more gaze shifts occurred in negative sessions.[12] At first sight, this may appear to be a contradiction to the significant less amount of gazes at the object for *success* than for *failure* scenes discussed above, but this is not the case: we evaluated whether or not at least one glance to an object (resp. other places) occurred in a particular interaction, whereas Lohse investigated the average duration of these glances. Furthermore, she compared positive and negative *sessions*, while we compared single *interactions*. Looking more while listening than while speaking as typical human behavior (please see Sec. 2.2.2) was not confirmed. Lohse suggested two plausible reasons for that: the subjects often looked to the side while listening (to aid concentration on what the robot was saying), and they frequently looked down when they put an object back to the table [333, pp. 117–119].

3.4.3. Facial Expressions

The facial muscle movements caused by the verbal reactions of the subjects constituted the by far most prominent type of facial expressions. The subjects talked in about 98% of all interactions (Tab. 3.2, rows 17 and 18). In *failure* scenes, the subjects very often verbally corrected the robot, whereby they used a pronounced speech highly significantly[13] more often than in *success*ful interactions (65.7% vs. 18.5%, Tab. 3.2, rows 19 and 20). The variance between different persons was very high (subjects 03, 06, 07 and 10 used pronounced speaking much less than the others). Nevertheless, pronounced speaking gives some useful indication for *failure*, but might be difficult to recognize by vision only (the observations discussed above are also largely based on audio).

Apart from talking, clearly visible facial expressions occurred comparatively seldom. Laughter occurred tendentially more often in *failure* scenes, but the difference to *success* scenes was not significant,[14] variance high again and three subjects showed the opposite pattern (Tab. 3.2, rows 21 and 22). In many cases, a laughter either was preceded by Biron cutting in during the subject was speaking (e.g. by uttering "That's interesting." or "I like it.") or it was due to amusement because of Biron's mistakes. A laughter that appeared to be related to joy about Biron's good performance was found only a few times. In a few places, a kind of affirmative expression—in most cases a perking up the eyebrows in a way—was found (Tab. 3.2, rows 23 and 24). The differences between *success* and *failure* were very small and not significant.[15] It did not become clear whether this expression actually is a means to affirm or support a (positive or negative) verbal answer, a sign for awaiting an answer, or something else. This seemed to be different for the individual participants, but was not investigated in greater detail (an affirmative or emphasizing function could be expected according to the studies of Grammer *et al.* as discussed in Sec. 2.2.3). In *failure* scenes, various "negative" facial expressions appeared (Tab. 3.2, row 25). These were typically a sort of screwing up of the eyes, often combined with frowning, but also some lip movements or a negatively perceived

[11] those sessions where Biron correctly classified the majority of objects
[12] those sessions where Biron misclassified the majority of objects
[13] $p < 0.001$ in both two-tailed t-test and Wilcoxon rank sum test
[14] $p > 0.1$ in both two-tailed t-test and Wilcoxon rank sum test
[15] $p > 0.9$ in both two-tailed t-test and Wilcoxon rank sum test

configuration of the face as a whole. Only in one instance, a negative expression was found in a *success*ful interaction: subject 07 reacted to Biron's correct classification as "book" by reading a few phrases from the book and frowned at some complicated mathematical term—a very untypical situation. Thus, these negative facial expressions are an indication for failure, but occurred only in 24.5% of the *failure* scenes (with a high variance). Moreover, almost half of these expressions were very weakly pronounced and difficult to recognize, for an automatic system probably even more than for humans (Tab. 3.2, row 26).

3.4.4. FCS Display in the Related Study of Barkhuysen *et al.* [18]

Barkhuysen *et al.* [18] also conducted a study where people displayed FCSs as reaction to a positive or negative interaction situation; their experiments will be discussed in Sec. 3.6.1. They found smiles, head movements, diverted gaze, frowning, eyebrow raising, "final mouth opening", repetitive head movements, and audiovisual hyperarticulation as frequently occurring cues. While the first seven are purely visual features, the analysis of the last one was largely based on audio, but Barkhuysen *et al.* [18] stated that this hyperarticulation was also visually cued, in accordance with the findings of Erickson *et al.* [146].[16] The "final mouth opening" cue is a silent opening of the mouth at the end of an evaluated video, where the person prepares for upcoming speech. This cue is "special" in some sense and might be specific for their scenario (please see Sec. 3.7).

The other cues are very similar to the FCSs we found in our study. Barkhuysen *et al.* [18] showed that the presence of these cues was perceived as indication for a negative interaction by most observers (except for final mouth oping which was positively rated), even more if these cues occurred in combination. However, their were in part large differences between the three experiments they performed (please see Sec. 3.6.1) and also varying degrees of significance. An analysis of the *actual* interaction situation (positive or negative) confirmed these findings from the analysis of the *perceived* situation for the most part, even though several relations were not significant any more (which is very likely due to the much small number of datapoints in this case). Head nods were an interesting exception, because they also occurred frequently in negative interactions, similarly, frowning was found several times also in positive interactions. Both issues were not only different from the observing subjects' perception, but also differed notably from the display of FCSs in our object-teaching study. Apart from this, the display of FCSs in their and our study appear to be fairly similar. Also their summarizing statement that typically a higher level of dynamic variation of facial features occurred in negative interactions is largely compatible with the display of facial expressions in our object-teaching study.

3.4.5. Conclusion

The participants of the object-teaching study did indeed frequently display FCSs during their interactions with Biron. For each category—head gestures, eye gaze, and facial expressions—behaviors that were more typical for either *success* or *failure* were identified. However, the variance between different persons was very high in most cases, the FCSs in part difficult to

[16] However, the evidence for that seemed to be anecdotal, as they wrote later: "In general, it would be interesting to redo the three experiments in a vision-only setting, to find out whether there are indeed visual correlates of hyperarticulation." [18, p. 28]

Figure 3.3.: Examples of various kinds of FCSs that occurred during the interactions in the object-teaching scenario. First row: Four frames of a head nod (subject 03). Second row: Four frames of a head shake (subject 07). The middle images are also examples for closing eyes in a *failure* scene, the outer images for looking at the robot. Third row (left to right): Averted gaze at beginning of scene, gazing at the object in a *failure* scene, looking down to the object table at the end of a *success*ful interaction, and gazing elsewhere during the interaction (subjects 04, 08, 11, and 10). Fourth row (left to right): very pronounced speaking, laughter, frowning and screwing up eyes (all three are reactions to *failure*), and affirmative perking up of eyebrows as reaction to *success* (subjects 09, 01, 02, and 06). Please refer to Sec. 3.4.

recognize, and the indicative power[17] of a signal often not strong enough to base a classification solely on this signal. Thus, we need to consider a combination of these three kinds of FCSs in order to tackle a reasonable classification into *success* and *failure*. Due to the large differences between the subjects, this combination most likely needs to be person-specific. The way the subjects displayed FCSs in the object-teaching study is very consistent with the conclusions we drew in Sec. 2.3.

3.5. Human Valence Recognition Performance

This section reports our investigation of the human capability to interpret the FCSs shown by the participants of the object-teaching study in terms of *valence*. The goal of this feedback interpretation study is to find out how good humans are in distinguishing *success* from *failure* scenes, depending on the available context information. The results shall also serve as a baseline for automatic recognition approaches. The next Sec. 3.5.1 describes the performed study, the following Sec. 3.5.2 its results. Finally, Sec. 3.5.3 draws conclusions for a valence-based automatic interpretation of these FCSs.

3.5.1. Procedure

We randomly selected 88 object-teaching scenes: 44 *success* and 44 *failure* scenes (four *success* and four *failure* scenes for each of the 11 subjects). These 88 videos were presented to 44 new subjects (15 female and 29 male, ranging from 16 to 70 years in age) who were not involved in the object-teaching user study. Their task was to decide by forced choice whether the displayed interaction situation was a *success* (Biron termed the object correctly) or a *failure* scene (Biron named the object incorrectly). To investigate the influence of the temporal and visual context on the valence recognition performance, we varied the amount of displayed context in two respects:

1. *temporal context:* showing the full video sequence versus showing the first half of it only
2. *visual context:* showing the full scene versus showing the face only

Combined with each other, this results in four different variants of each video. The bounding boxes of the faces were located with the automatic face detection approach described in Sec. 4.1.6.[18] All videos were presented without sound, because hearing the subject's verbal answers would make the distinction between *success* and *failure* trivial in most cases. To allow for a suitable evaluation of the four context conditions, the video sequences were distributed over the 44 subjects such that the following conditions were met:

- Each subject saw each video sequence under one context condition only. To avoid priming effects, we did not show the same scene twice (in different context variants) to the same person.
- Each subject saw all 88 videos (and thus four *success* and four *failure* scenes for each of the 11 persons from the object-teaching study) in randomized order.

[17] In several cases, the frequency differences of a specific FCS between *success* and *failure* were, although significant, not distinct enough.

[18] The automatic face detection led to a kind of "glint" around the faces (as the face size varies somewhat) in some cases, also in a few cases the face detection got lost for a few frames. Videos where the face detection was too poor were rejected beforehand.

	all videos			*success* videos			*failure* videos		
	mean	SD over		mean	SD over		mean	SD over	
context variant	**acc.**	**sub.**	**vid.**	**acc.**	**sub.**	**vid.**	**acc.**	**sub.**	**vid.**
all context variants	79.1	8.2	17.9	75.8	11.9	19.4	82.4	12.0	15.8
full-scene/full-time	83.4	12.8	18.1	80.2	16.8	21.1	86.6	16.2	14.1
full-scene/first-half	78.2	8.1	24.0	75.0	12.6	27.2	81.4	15.3	20.1
only-face/full-time	82.0	11.1	19.1	78.1	16.3	21.2	86.0	13.5	16.1
only-face/first-half	72.8	9.9	23.9	69.8	15.9	25.8	75.8	15.9	21.7

Table 3.3.: Mean classification accuracy ("acc.", percentage) and standard deviation (SD) for all videos, only *success*, and only *failure* videos, each for different context conditions as well as the average of all four context conditions. In each case, there are two different SD values: one for the distribution over the observing subjects ("sub."), and one for the distribution over the judged videos ("vid."). Please refer to Sec. 3.5.2.

- Each subject saw exactly 22 videos in each of the four context variants in randomized order (11 *success* and 11 *failure* scenes)
- Summed up over all 44 subjects, each video was seen by 11 subjects in each of the four context variants.

Figure 3.4 depicts example images from two of these video sequences. The subjects made a forced decision between *success* and *failure* for very video, skipping videos was not allowed. They could watch each video as often as they liked, but were asked to make their (final) decision before they moved on to the next video.[19]

3.5.2. Results

Averaged over all context variants, the subjects of the feedback interpretation study were able to classify the videos with 79.1% classification accuracy. We did not observe differences between female and male observers, the classification rate was 79.1% for both. Table 3.3 lists the mean classification accuracy and standard deviation for the different context conditions. There were big differences between the subjects, ranging from 90% to 59% accuracy on average. The visual context helped in the classification, as the performance was better for "full-scene" videos compared to "only-face" videos, significantly for "first-half" videos ($p < 0.01$) and very slightly only (not significantly) for "full-time" videos ($p > 0.4$).[20] The temporal context was more important, as the subjects performed better for "full-time" videos compared to "first-half" videos, and the difference was greater than for the visual context. This effect was significant for both "full-scene" ($p < 0.03$) and "only-face" ($p < 0.001$) videos. On average, *failure* videos were recognized significantly better than *success* videos ($p < 0.011$). For both classes individually, the variance was higher than the total variance over all videos, because most subjects (26) were better in classifying *failure* videos than in classifying *success* videos, but for some subjects (12) the opposite was the case (six subjects performed equally well in either case); thus these differences partially level in the overall statistics over both classes.

[19] Please refer to App. A.4 for a copy of the instructions the subjects received.

[20] Throughout this section, the significance of the discussed results was tested by both a two-tailed t-test and a Wilcoxon rank sum test, for reasons similar to those outlined in footnote 10 of this chapter. The given p-value relations hold for both tests in each case.

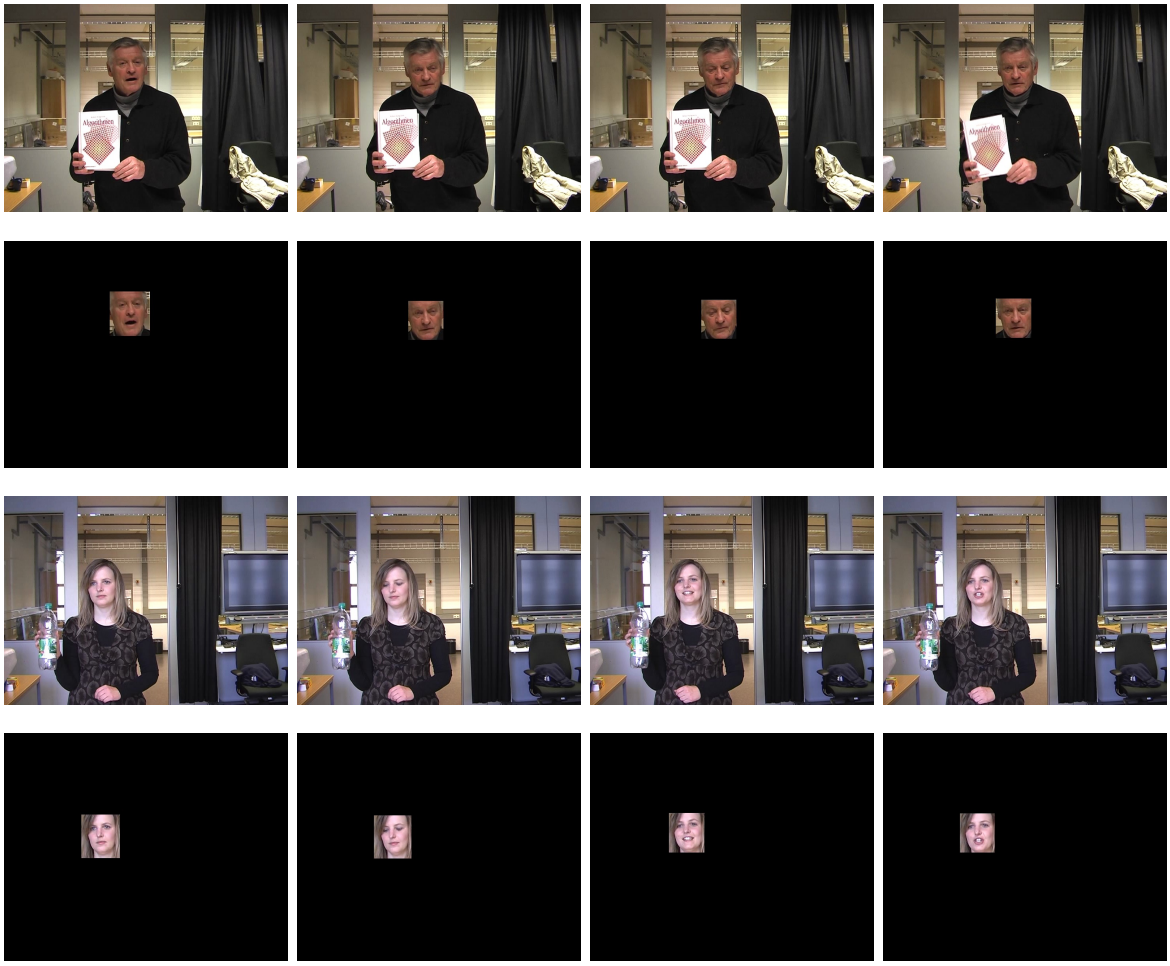


Figure 3.4.: Example images from the video sequences the subjects of the feedback interpretation study judged. First row: Four frames of a *success* video, showing the full scene (subject 05). Second row: Four frames of the same video, showing only the face. Third row: Four frames of a *failure* video, showing the full scene (subject 11). Fourth row: Four frames of the same video, showing only the face. In each case, the four frames are approximately equidistantly distributed over the video sequence. Thus, the last two frames are always part of the full-time video sequence only, but are not contained in the videos that show the first half only. Please refer to Sec. 3.5.1.

The variance between different videos was even higher than the variance between subjects: some videos were correctly classified in almost every case, whereas some other videos were systematically misclassified. The most recognized video showed a clearly visible nodding, the second most recognized video contained clear signs that the subject was perplexed. In the most poorly recognized video, the subject spoke a lot without clear affirmation, which was misinterpreted by most subjects as correcting the robot. The subject shown in the second most poorly recognized video displayed hardly any prominent FCS at all. The variance for *success* videos was higher than for *failure* videos, which is consistent with the observation that *failure* videos were easier to classify on average, but there were also some *success* videos that

sub.	full-scene / full-time						only-face / full-time					
	all		*success*		*failure*		all		*success*		*failure*	
	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**
01	84	25	98	5	68	28	82	19	91	7	73	23
02	82	18	75	25	89	5	75	18	66	19	84	14
03	90	11	84	14	95	5	85	19	84	20	86	22
04	90	12	84	14	95	9	92	6	89	5	95	5
05	74	25	68	28	80	24	68	19	61	16	75	21
06	82	7	84	9	80	5	73	19	70	25	75	14
07	92	8	98	5	86	5	94	10	91	13	98	5
08	75	27	61	32	89	14	67	28	52	30	82	20
09	68	21	50	12	86	5	78	21	66	20	91	13
10	91	10	86	12	95	5	95	7	95	9	95	5
11	91	10	93	9	89	11	92	10	93	9	91	13

Table 3.4.: Mean classification accuracy ($\overline{\mathrm{m}}$, percentage) and standard deviation (SD) for the videos of the 11 subjects of the object-teaching study in the "full-time" context condition. The left block shows the results for all videos, only *success*, and only *failure* videos, each for the "full-scene" context condition; the right block shows the respective results for the "only-face" context condition. Please refer to Sec. 3.5.2.

were correctly classified in almost every case. Tables 3.4 and 3.5 show the mean classification accuracy and standard deviation itemized for the 11 subjects of the object-teaching study. As the detailed results listed in these tables are based on four *success* and four *failure* videos per subject only (judged by 11 observers in each of the four context conditions), they should not be over-interpreted. Nevertheless, it can be seen that there are large differences between the subjects, not only regarding the mean classification accuracy, but also regarding the standard deviation and thus the single videos of a subject: for some subjects, the "difficulty level" of their videos seems to be similar, whereas it apparently differs notably for other subjects. Figures 3.5, 3.6 and 3.7 illustrate the distribution of the achieved classification accuracy over the observing subjects and also over the judged videos in detail.

3.5.3. Conclusion

The subjects of this feedback interpretation study were able to distinguish *success* from *failure* videos with recognition performances between 73% and 83% on average, but there were in part large differences depending on:

- the subject that interacted with the robot
- the concrete interaction video of this subject
- the observing resp. judging subject
- the amount of displayed temporal and visual context

These results have important implications for an automatic recognition of FCSs in the object-teaching scenario. The visual context is not that important: given a sufficient length of the video ("full-time" condition), there was no significant difference between the recognition performance for "full-scene" and "only-face" videos. Thus, automatic recognition approaches can reasonably be restricted to process only the face of the interacting subject, it is not

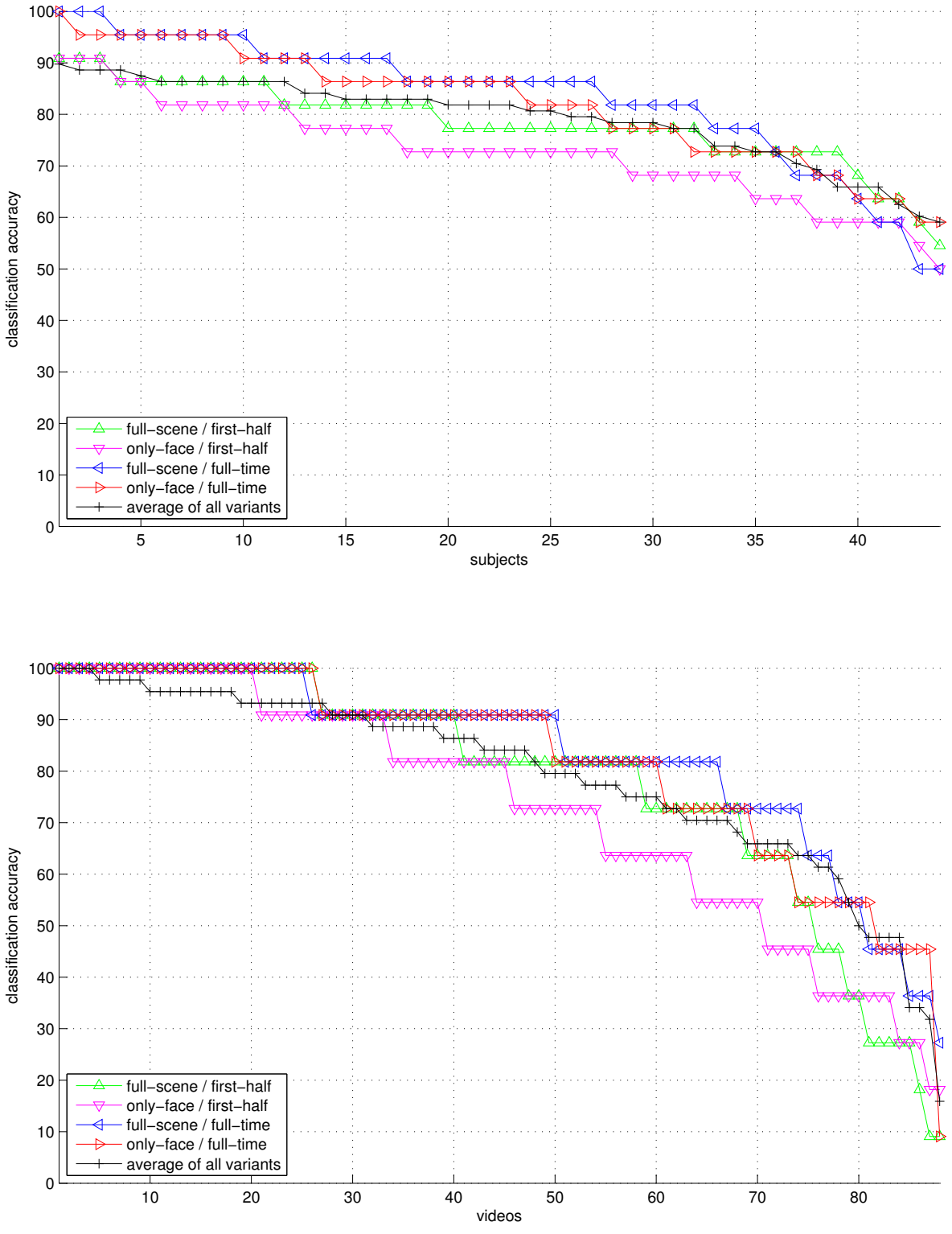


Figure 3.5.: The classification accuracies for all 88 videos for the four context variants and the average of them, respectively. Top: Classification accuracy distribution over the observing subjects (sorted for each context variant). Bottom: Classification accuracy distribution over the videos (sorted for each context variant). Please refer to Sec. 3.5.2.

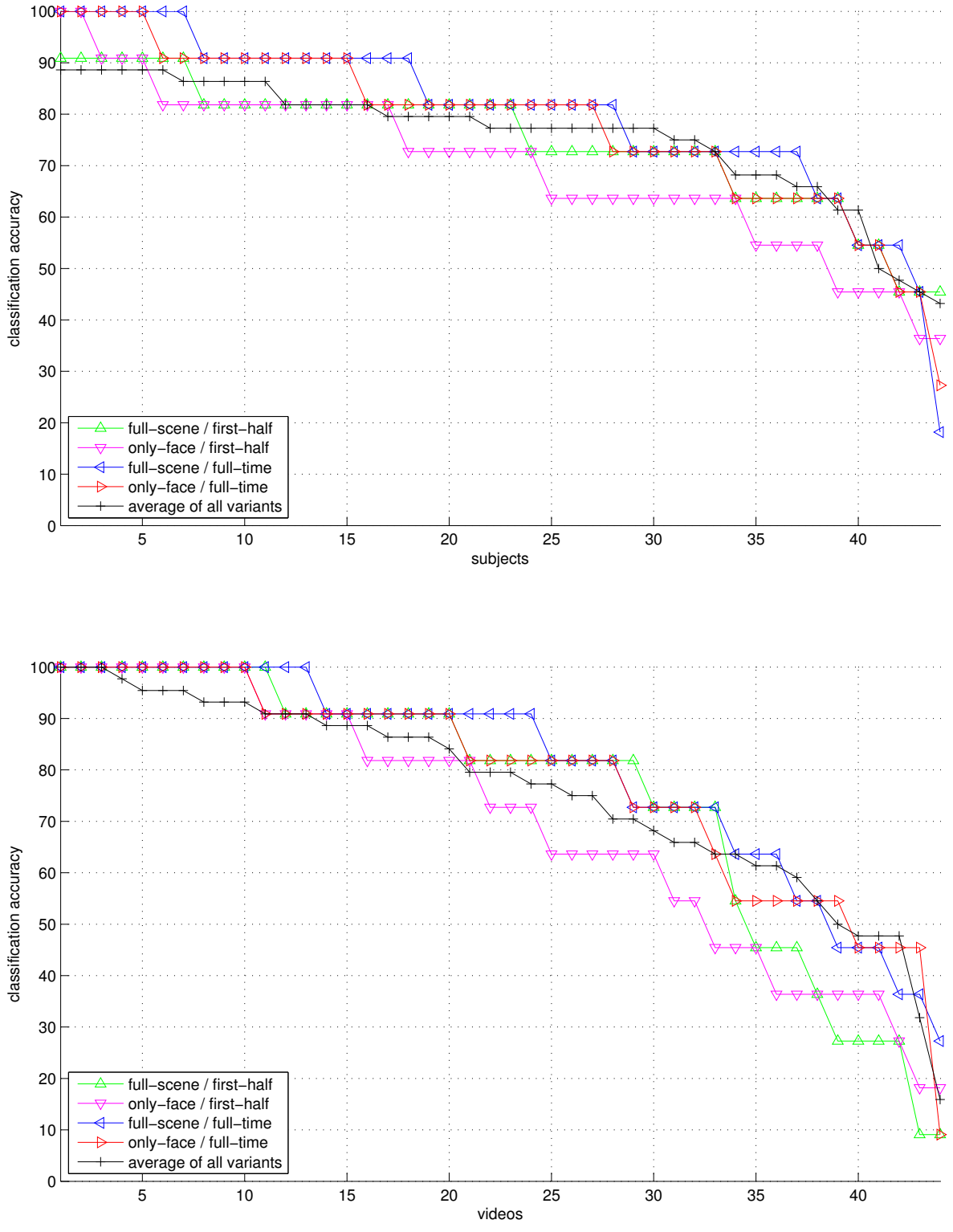


Figure 3.6.: The classification accuracies for the 44 *success* videos for the four context variants and the average of them, respectively. Top: Classification accuracy distribution over the observing subjects (sorted for each context variant). Bottom: Classification accuracy distribution over the *success* videos (sorted for each context variant). Please refer to Sec. 3.5.2.

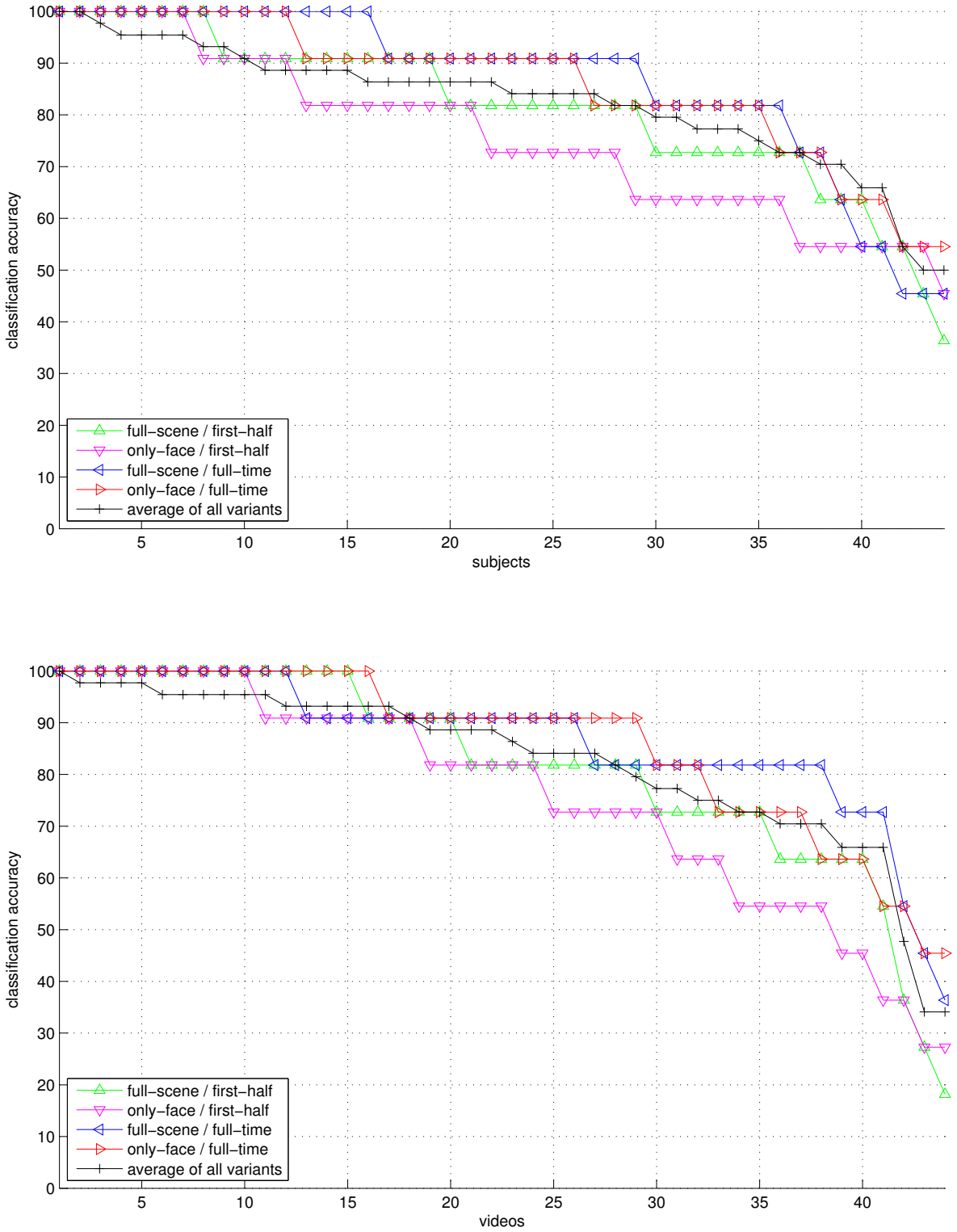


Figure 3.7.: The classification accuracies for the 44 *failure* videos for the four context variants and the average of them, respectively. Top: Classification accuracy distribution over the observing subjects (sorted for each context variant). Bottom: Classification accuracy distribution over the *failure* videos (sorted for each context variant). Please refer to Sec. 3.5.2.

sub.	full-scene / first-half						only-face / first-half					
	all		*success*		*failure*		all		*success*		*failure*	
	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**	$\overline{\mathbf{m}}$	**SD**
01	73	28	89	9	57	33	69	27	84	11	55	32
02	86	10	86	12	86	9	77	16	70	9	84	20
03	88	16	89	11	86	22	92	16	98	5	86	22
04	87	9	89	5	89	14	85	14	86	17	84	11
05	67	39	57	42	77	40	63	30	61	31	64	33
06	81	20	82	26	80	16	66	19	64	27	68	12
07	98	6	95	9	100	0	86	22	82	31	91	13
08	72	28	66	40	77	12	66	26	52	30	80	14
09	55	22	36	10	73	13	49	19	41	9	57	24
10	77	22	68	28	86	12	81	20	73	23	89	14
11	76	25	68	34	84	11	67	25	57	33	77	12

Table 3.5.: Mean classification accuracy ($\overline{\mathrm{m}}$, percentage) and standard deviation (SD) for the videos of the 11 subjects of the object-teaching study in the "first-half" context condition. The left block shows the results for all videos, only *success*, and only *failure* videos, each for the "full-scene" context condition; the right block shows the respective results for the "only-face" context condition. Please refer to Sec. 3.5.2.

mandatory to consider the visual context around. On the contrary, the temporal context is important, as "full-time" videos were recognized significantly better than "first-half" videos. The variance of the classification accuracy was very high in most cases, which is not surprising due to the complex nature of FCSs. Thus, a high variance can also be expected for automatic recognition techniques.

3.6. Related Works

There exist a large number of other databases of people showing various facial displays, especially focusing on facial expressions. Many of them show posed facial expressions, most often in terms of basic emotions (and several additional categories in some cases), for instance the Cohn-Kanade database [266], the FABO database [203], the DaFEx database [24], and the databases of Chen [76] and Yin *et al.* [541]. Several researchers created databases of spontaneously occurring facial expressions. Sebe *et al.* [451] used a video kiosk setting to record people watching movie trailers that are expected to induce basic emotions, O'Toole *et al.* [387] and also McDuff *et al.* [352] used a similar approach. Bartlett *et al.* [23] introduced a video database where people displayed spontaneous FCSs while they are engaged in political or social discussions, where the FACS codings [139] of the faces are provided. A database designed to investigate differences in the listening behaviors of people while a short video is retold or a recipe is described to them was presented by de Kok and Heylen [107].

The Belfast database [121] contains television clips and interviews were people display authentic FCSs (or at least appear to do so [121, p. 49]) where a dimensional labeling in the activation-evaluation space was produced by observers using the "Feeltrace" tool [96]. This kind of labeling was also used to annotate the SAL database (which is part of the HUMAINE database [122]). It shows users interacting with artificial listeners of different characters which

tried to induce positive or negative emotions. The system was designed based on a comprehensive evaluation of respective psychological work [97], following the approach of Weizenbaum [516]: a computer program tries to give the impression of a "real" conversation and understanding, while in fact it just applies a set of rules and transformations of the input (the text spoken by the user) to produce its answers. The users showed spontaneous facial displays, however, they were aware that the interaction is about their emotions ("emotional gym", [166, p. 397]), which might have had some influence. The evaluation dimension in these two databases can be viewed as positive or negative valence.

For all the databases discussed above, the ground truth labels are defined based on the visual appearance of the face, usually by self-report of the subjects (e.g. [451, 352]), observing judges (e.g. [387]) or direct FACS coding (e.g. [23]), in contrast to the definition in terms of the interaction situation in our database (please see Sec. 3.3). An exception is the work of Barkhuysen *et al.* [18] who used a similar approach; we compare their study to ours in Sec. 3.6.1.

There are a few studies who investigated the human recognition performance for valence (resp. positive and negative ratings in general). Simon *et al.* [459] investigated observer judgments for one-second videos showing facial expressions of pain and basic emotions. The observers were able to judge these facial displays in terms of intensity, arousal, and valence with high sensitivity and specificity. Gilbert *et al.* [188] evaluated the performance of observers judging videos of people reacting to unpleasant, neutral, or pleasant odors. The classification accuracy was 76% for subjects posing facial displays to real or imagined odors, but dropped to 37% when those displays were spontaneous. Krahmer et al. [293] showed that people can correctly classify disconfirmation fragments of dialogs as positive or negative communication signals. They used audio recordings instead of facial displays.

3.6.1. The Study of Barkhuysen *et al.* [18]

Barkhuysen *et al.* [18] conducted three experiments where 66 subjects (20 male and 46 female) watched film fragments of speakers interacting with an oral train timetable dialog system. The subjects decided whether or not there was a communication problem present in the shown situation. In each experiment, the ground truth data was defined in terms of the interaction situation, similar to our approach. The types of displayed FCSs have already been discussed in Sec. 3.4.4, hence we focus on the recognition performance here. Per speaker, two problematic and two non-problematic videos were shown in each experiment.

In the first experiment, the subjects observed nine silent speakers listening to a confirmation question of the system, which revealed whether the system's recognition of the desired travel destination was correct or not. About 75% of the subjects classified the videos correctly, and about 70% of the videos were significantly classified correctly. Likewise to our study, the standard deviation over speakers was very high and systematic misclassifications occurred for some videos.

In the second experiment, the subjects watched videos of seven speakers saying "no", either in response to a yes-no question (non-problematic) or to indicate a misrecognition of the system (problematic). In this case, the classification accuracy was only slightly above chance level. Systematic misclassifications were relatively frequent, due to atypical FCSs or hardly any recognizable facial display shown by some speakers. Thus, this recognition task seemed

to be very hard, perhaps partially due to the short duration of the video sequences. Again, Barkhuysen *et al.* [18] observed large differences between different speakers.

In the third experiment, eight speakers uttered a destination, either in answering a question about the travel destination (non-problematic) or to correct a misinterpretation of their previous answers (problematic). About two thirds of the subjects classified these videos correctly, and most of the videos were significantly classified correctly. Again, the variance between different speakers resp. videos was very high. Significant misclassifications were mostly due to atypical behavior, for example hyperarticulation in non-problematic situations.

Overall, the results of these experiments and our study match fairly well. The partially higher classification accuracies in our study might be due to the different settings. In our object-teaching scenario, the videos seem to contain more "implicit" context that could help the interpretation of the observing subjects, especially regarding eye gaze: while there were at least two important gaze targets in our study—the robot and the object table—there was only one relevant gaze target (the camera) in the study of Barkhuysen *et al.* [18]. (Apart from diverted gaze which occured in both studies.) Indeed, when asked about the features they (believed to) have used to classify the videos,[21] some subjects of our study mentioned aside from head gestures, "lipreading", and facial expressions also some "implicit" contextual features: whether the person seemed to put down the object at the end of the video (thus shifting gaze from the robot to the table), and also the length of the sequence respectively how much the person was talking.

In spite of the similarities of the experiments, there are also some important differences. Whereas Barkhuysen and her colleagues varied the shown video sequences, we used the same video sequences and varied the amount of displayed visual and temporal context. They presented the videos with sound, whereas we removed the sound from all videos. Barkhuysen *et al.* [18] investigated differences between the videos respectively persons shown in the videos, we additionally reported about differences in the recognition performance of the observing subjects. Furthermore, the task of the people shown in the videos is different. This has some influence on the displayed FCSs, which is briefly discussed in the next section.

3.7. The Interaction Context

The discussion of FCSs in Sec. 2.2 shows that display and meaning of FCSs heavily depend on the interaction context in various ways. This context, however, is difficult to define. Commonly, this term is used as a very general, often vague concept, as Reich [414] pointed out: "The blunt truth is that social scientists have tended to treat context as a residual category that encompasses 'everything else' with regard to the object of interest (a communicative message, a social situation, and so on)." [414, p. 48]. This is not necessarily a negative point, because this view is very reasonable for many basic investigations of an "object of interest", where everything that is not explicitly considered, but might have an influence, is regarded as "context". However, when the goal is to equip an artificial agent (e.g. robot) with the capabilities to perform sophisticated interpretations of human FCSs in a wide range of situations, a concrete model of this "everything else" needs to be developed.

There are different perspectives on how such a model might be organized on top level. Reich [414] suggested three basic components artificial agents should consider when aiming to

[21] Please see App. A.5.

interpret communicative signals of humans (e.g. pointing gestures): action affordances, tool-mediated causal relationships, and social activities. His perspective appears to be focused on actions the human might want the artificial agent to perform. Ter Maat and Heylen [481] grouped contextual elements into three basic categories: parameters of the signal (e.g. speed of a head nod), constraining elements (e.g. greeting gesture only at beginning of interaction), and pointer elements (indicators for meaning disambiguation, e.g. co-occurrence of two signals). They further emphasized that the border between signal and context is often unclear and several appropriate ways to draw this line exit (e.g. a smile might occur in the context of a head nod, or the smile and the head nod together might form the signal). This contrasts Reich's view who regarded signals as "easily identifiable" [414, p. 50]. Several researchers investigated context from a linguistic perspective, concentrating on the context needed to understand and interpret speech or written text (e.g. [340, 206, 4, 499]). This indirectly also affects FCSs, because the speech accompanying them might be important for their interpretation. Ekman and Friesen [140] discussed in detail context conditions for the usage of nonverbal signals. Overall, a comprehensive and generally useful conceptualization of context is very complex and not achieved to date; many of the researchers cited above emphasized the need for further investigations. Nevertheless, what is clear is the large, non-negligible influence of the context.

In Sec. 2.3, we concluded that a general purpose interpretation of FCSs by robots is unlikely to be realizable in the near future, due to the complexity and context-dependence of FCSs. We further suggested a pragmatic simplification and focusing on different, specific interaction scenarios, like the object-teaching scenario we investigated. This "pragmatic simplification" also applies to the context, where we take the following view:

- The whole facial display is considered as *signal* (and not as *context*) in any case, in accordance with our definition of FCSs in Sec. 2.2. Thus, we do not distinguish between components of a facial display that are signals and others being context as ter Maat and Heylen [481] discussed. This also means that the parameters of a signal (in the sense of ter Maat and Heylen [481]) are always part of the *signal*, not the *context*. (Hence, two head nods of very different speed or amplitude would be viewed as two signals rather than one with different parameters in two cases.) This avoids the introduction of a suitable FCS parameter space and the need to determine these parameters (both problems might be challenging) at the cost of increasing the set of possible FCSs.
- Regarding the categories of interpretation, we already argued in Sec. 2.3 in favor of broader clusters like *positive* vs. *negative* (in case of our object-teaching scenario *success* vs. *failure*) because we expect them to generalize to a wider variety of contexts than finegrained categories. Valence, in the broad sense of *positive* vs. *negative*, appears to be a very general concept. This does not necessarily mean that the actual facial displays subsumed by these two categories are similar in appearance for two different contexts, but that *valence* as a general concept for interpretation is suitable in both cases. Nevertheless, this does not preclude using additional or other classes for certain interaction scenarios, which, however, might then be specific for these scenarios.
- We identify a *context* with a certain *interaction scenario*. This means, the object-teaching scenario constitutes one context. Others might be given by a "showing the robot around an apartment" scenario, a "collaboratively arrange furniture" scenario, or a "prepare and serve meal" scenario, for instance. The distinction between scenarios needs to be done carefully, especially regarding the decision whether two scenarios might

in fact be the same scenario. For instance, teaching objects to a robot in a laboratory while another person (a researcher) is present and the interaction is recorded on video—as it was the case for the object-teaching study investigated in this work—is likely to constitute another context than teaching objects alone at home (this can be expected according to the human audience effect studies discussed in Sec. 2.2.3). Also changes to the setting (e.g. placing the objects on a shelf instead a table) might influence the displayed FCSs (please see below). We suggest to treat those situations as different scenarios and possibly fuse them after an inspection revealed that the actually occurring FCSs are basically the same.

- When data is available from two or more scenarios, the FCSs that occurred there can be evaluated similarly to the evaluation presented in Sec. 3.4. This evaluation shall reveal which FCSs are specific to one scenario and which are applicable for several scenarios. More generally, for each type of FCS, a set of scenarios (resp. contexts) where this FCS—together with its interpretation in this scenario—is applicable is to be maintained. This information can then be incorporated in a robot system. During its interactions with humans, the robot needs to know to which scenario the current interaction belongs, of course.[22]

We investigated the object-teaching scenario in detail, but did not perform the evaluation described in the last item for a second scenario. Nevertheless, we can conjecture which FCSs might be specific to this scenario based on our experiences. This especially appears to be the case for some aspects of the gazing behavior: at the end of *success* scenes, the subjects often put the object back on the table and looked down while doing so. They would probably act differently if the objects were not placed on a table, but on a shelf at eye level, resp. in general for tasks where putting something down is not common. The subjects also gazed significantly more often at the object in *failure* scenes. This might be interpreted as a signal for concentration resp. thinking about how to resolve the current failure situation (by looking at an essentially involved object, please see the discussion of gaze and attention in Sec. 2.2.2). In tasks where no relevant, manipulable objects are present, this signal might be replaced by averted gaze (please see the discussion of Kendon's work in Sec. 2.2.2). Most of the remaining FCSs, in particular head nods for *success* and head shakes, pronounced talking, and "negative" facial expressions for *failure*, are expected to be relevant for a wider range of scenarios.

This conjecture is supported by the FCS evaluation Barkhuysen *et al.* [18] did for their study. As discussed in Sec. 3.4.4, most of the signals they found match signals in our study fairly well, which suggest a certain generalizability. An exception is the "final mouth opening" cue, which seems to be specific to their scenario in the first place. However, it occurred when "a speaker silently opened his mouth at the end of the video film to prepare for upcoming speech" [18, p. 350]. Thus, it might be special because of the signal segmentation Barkhuysen *et al.* [18] used: in case one would consider additional parts of those videos in another evaluation, it might turn out that these "final mouth opening" cues form the beginning of hyperarticulated speech, which also occurred in our study, thus this cue would be more general in that case. Since this is speculative, the specificity or generality of this cue cannot be conclusively determined here, though.

[22] One possibility is to try to get this information from the human interaction partner by speech recognition (e.g. recognizing "I want to teach you some objects."). A "general" fallback scenario (with a small set of FCSs that are found or assumed to be applicable for a wide range of contexts) could be used if no specific scenario can be determined.

4. Automatic Recognition of Facial Communicative Signals

The wonderful thing about standards is that
there are so many of them to choose from.
— attributed to Grace Hopper, Andrew S. Tanenbaum, and others

This chapter provides an overview of state of the art approaches for an automatic FCS recognition. At first, face detection methods are discussed in Sec. 4.1, where we considered two approaches in greater detail in Sec. 4.1.5 and 4.1.6, because we use these approaches in our investigations in the next two chapters: the boosting approach of Viola and Jones [504] and the *Encara* approach of Castrillón *et al.* [64].

Subsequently, automatic classification approaches for the recognition of head gestures, eye gaze, and facial expressions are considered in Sec. 4.2, 4.3, and 4.4, respectively. Finally, Sec. 4.5 discusses feature extraction methods we apply in the following chapters in some more detail, in particular active appearance models in Sec. 4.5.1, but also (more briefly) constrained local models in Sec. 4.5.2 and Gabor energy filters in Sec. 4.5.3.

4.1. Face Detection

The automatic detection of human faces in images or videos has been intensively investigated in the recent 25 years. Several researchers compiled comprehensive surveys of the proposed approaches [435, 74, 221, 534, 110, 551]. In our discussion here, we follow the categorization of Yang *et al.* [534] who distinguished four basic approaches which we briefly discuss in the following sections 4.1.1 to 4.1.4: knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods. However, this is not a firm classification, because many approaches could be ranged in more than one category [534], and other categorizations according to different criteria are possible, of course (e.g. [221]). Section 4.1.5 describes the face detection technique of Viola and Jones [504] in greater detail as this is a *de facto* standard nowadays. The last Sec. 4.1.6 considers an extension of this approach, the face detection and tracking technique developed by Castrillón *et al.* [64], which we used in our work reported in the subsequent chapters 5 and 6.

4.1.1. Knowledge-Based Methods

Knowledge-based face detection methods are top-down approaches that directly incorporate human knowledge of the typical structure of a human face into algorithmic rules. Usually, specific relations of prominent face characteristics are modeled, e.g. the presence of two

symmetric eyes in certain distances to nose and mouth [534]. A practical challenge of these approaches is the appropriate definition of the rules in order to achieve high true positive and at the same time low false positive detection rates for a wide range of faces in different poses [534]. We briefly discuss a few representatives of this category below.

Yang and Huang [531] presented a system that processes multiresolution images where face candidates are searched with a scanning window at low resolutions; found candidates are then further processed at finer resolutions. Faces are discriminated from non-faces by three levels of rules. At the first level, rules about the expected arrangement of regions of nearly uniform intensity are applied. The second level searches for local minima within the candidate face regions that might correspond to eyes, nose and mouth. The surviving candidates are verified at the third level where an edge detection is performed and false positives are rejected based on atypical edge configurations in eye and mouth regions. Kotropoulos and Pitas [292] proposed an improvement of this method. To speed up the face candidate detection, they use horizontal and vertical profiles [265] and search them for local extremes that correspond to the left and right side of the head (horizontal profile) and positions of eyes, mouth lips, and nose tip (vertical profile).

Hsu *et al.* [229] developed a method that finds face candidates as elliptical skin-colored regions. The candidates are verified by the detection of eyes, mouth, and face boundary based on typical color and intensity relations and the geometric configuration of the facial features. Fei and Qiang's approach [159] also detects skin-colored regions first and applies at set of heuristics about the face dimensions and the positions of facial features and their relations to local intensity minima afterwards. Jeng *et al.* [252] used a geometric face model that examines the vertical distances of eyebrows, eyes, nose, and mouth to find faces. The face detection approach of Castrillón *et al.* [64], which we discuss in Sec. 4.1.6 in greater detail, includes prior knowledge about a typical face configuration as well.

4.1.2. Feature Invariant Approaches

Feature invariant approaches [534] try to detect features that are to a reasonable degree invariant to face pose, viewing angle, or lighting conditions. Thus, these features can be used to locate faces under a wide range of conditions. However, the most serious problem for these techniques is that the invariance assumptions might be only partly fulfilled for real world data due to illumination, shadows, noise, and occlusion [534].

A great many methods detect individual facial features such as eyebrows, eyes, nose, mouth, or hair-line, usually based on edge detectors, and use various perceptual grouping techniques to infer the presence of a face from the detected features [534]. Sirohey [463] applied a standard edge detection technique [60] and heuristically grouped the found edges to test for faceness by means of ellipse fitting. Chetverikov and Lerch [77] detected dark and light blobs in a certain configuration and used "streaks" of edges with similar orientation to find faces. Yow and Cipolla [546] presented a probabilistic framework that performed perceptual grouping using evidence propagation in a Bayesian network based on geometrical, spatial, and intensity relations. Leung *et al.* [317] used a probabilistic model as well, but distinguished faces from other objects by graph-matching techniques based on a statistical model of mutual distances between facial features. They applied Gaussian filters for feature detection, a technique that was also used by others [360, 475].

Another common approach is to find face candidates by detecting skin-colored blobs [534]. There is evidence that the chrominance of human skin is stable enough to be used for skin-color detection [194, 532], despite possibly large intensity variations resulting from different lighting conditions. Therefore, skin-color detection approaches often use a color space that normalizes the intensity and thus tries to focus solely on the chrominance by removing the brightness-dependency [525]. However, there is also some evidence that the variations induced by light source and camera type require an explicit modeling to be appropriately normalized [471, 470, 177]. Although a carefully chosen range of color values was also used directly for skin-color-based face detection [69], more commonly derived representations, for instance based on histograms [441, 102, 492, 553, 331] or Gaussian distribution models [283, 482, 58, 354, 267], were utilized. In the latter case, often a mixture model was used because the skin colors of different ethnic groups yield a multimodal distribution [534]. However, according to the comprehensive study of Jones and Rehg [261], histogram-based models are superior to mixture models in terms of both accuracy and computational cost. Several approaches try to compensate for varying lighting conditions [229] or to adapt to them [354, 331]. Usually, face candidates found as skin-colored blobs are further processed by other means to confirm or reject them as faces [466, 533, 553, 65, 6].

4.1.3. Template Matching Methods

Template matching methods [534] use a stored face pattern that is compared against subwindows of the input image, usually by means of correlation methods. Kwon and da Vitoria Lobo [300] detected the face outline with edge features and used a geometric face template to verify the found candidates by minimizing a fitting error function with gradient descent. Similarly, the approach of Craw *et al.* [98] first detects edges using the Sobel operator and applies a template of the head outline. Subsequently, templates of lips, eyebrows, and eyes are used to find facial features to confirm the face presence. In later work, Craw *et al.* [99] refined their approach. They fitted a polygonal template of the head to an image region by simulated annealing [286] and used a set of feature experts to verify and refine the face location. Tsukamoto1 *et al.* [493] introduced a "Qualitative Model for Face (QMF)" [493, p. 754] where a face is modeled by several template blocks. Each block is associated with specific "lightness" and "edgeness" values that are used to compute the "faceness" of matching image parts which needs to exceed a threshold for a face to be detected. Samal and Iyengar [436] combined silhouette templates, principal component analysis (PCA) [285], and generalized Hough transformation [126, 227] to find faces. More recently, Jin *et al.* [254, 255] performed face candidate detection based on skin-colored regions and verified the faces by normalization and matching to a template of the whole face.

A general issue with template matching approaches is the dependence of the templates on scale, pose, and shape of the face [534]. Various methods to deal with this problem have been investigated. Craw *et al.* [99] used random transformations (regarding scale and orientation) of a head template to find the best fit to the input image. Deformable templates, a similar technique, were introduced by Yuille *et al.* [547]. The exact form of such a template is defined by several parameters which are optimized by minimizing a complex energy function with gradient descent. Huang [234] applied multiple rotated face templates to account for different face orientations.

A large number of flexible template matching models is based on *snakes* [274] or similar approaches, where an object contour model is fitted to an image region by energy minimization,

considering model parameters that describe the allowed deformations and image features such as lines and edges. Kwon and da Vitoria Lobo [300] adapted this technique in their *snakelet*-based approach to detect the face contour, Lam and Yan [302] developed a similar method. Snakes also inspired Cootes *et al.* [86, 91] who developed the *active shape model* (ASM) where the allowed deformations are required to be compatible with hand-annotated example deformations given by a training set. Though originally used to model and locate other objects (resistors, hearts, hands), ASMs were successfully applied to face detection by Lanitis *et al.* [313] and also by Cootes and Taylor [92]. An additional advantage of these methods is that the model parameters describe the shape and pixel intensities of a fitted face image and can thus be used as feature vectors for subsequent classification tasks.

4.1.4. Appearance-Based Methods

Appearance-based methods [534] are similar to template matching techniques, with the key difference that the templates are not designed exploiting human expert knowledge about the structure of faces, but they are gained from example images of faces and non-faces using machine learning methods. Most recent face detection approaches are at least in part appearance-based [551], as superior performance has been demonstrated for approaches of this category (e.g. [504]).

Turk and Pentland [496] presented a face recognition and detection approach based on *eigenfaces*, which are the eigenvectors gained from a PCA applied to a set of face images, thus the eigenfaces span a linear subspace of face images. The face detection is performed by the projection of candidate regions in the input image to this subspace: as face images are modeled by this subspace, they are expected to be affected little by this projection, in contrast to images of other objects that should not have a good representation in this subspace. Thus, the distance between the candidate region and the face subspace is used for the distinction between faces and non-faces. An advantage of this method is that it does not require a training set of non-face images.

However, most approaches rely on such a training set of non-face images to improve the discrimination boundary of the utilized classifier. Various classification techniques have been proposed for this discrimination of faces and non-faces. Sung and Poggio [473] used a multilayer perceptron (MLP) [213] to classify feature vectors representing differences of face candidate regions to several face and non-face clusters. Rowley *et al.* [423] also utilized MLPs, accompanied by arbitration methods to combine the outputs of several networks in order to improve the overall accuracy. Schneiderman and Kanade [444] used a Bayes classifier, Pham *et al.* [401] Bayesian network classifiers, Lin *et al.* [329] a probabilistic decision-based neural network, Samaria and Young [437] and also Nefian and Hayes [370] hidden Markov models (HMMs) [163], Osuna *et al.* [386] and also Castrillón *et al.* [64] support vector machines (SVMs) [445], and Huang *et al.* [233] decision trees [408]; please refer to the survey of Yang *et al.* [534] for a more detailed discussion. Other appearance-based face detection methods are based on entropy [5], kernel PCA [319], or Markov random fields (MRFs) [318, 284], for instance.

A disadvantage of several appearance-based approaches is that they are computationally expensive [221]. Viola and Jones [503, 504] presented a face detection approach based on boosted classifiers that efficiently addresses this issue; please refer to next section. According to the recent review of Zhang and Zhang [551], an adaptation of the face classifier to new environments

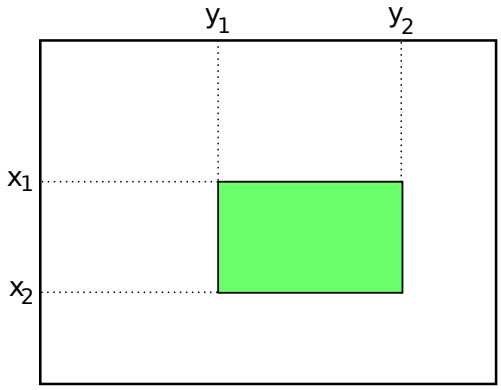


Figure 4.1.: Example illustration of image energy computation using the integral image ii [504]. The energy E of the green rectangular region can be computed in constant time: $E = ii(x_1, y_1) + ii(x_2, y_2) - ii(x_1, y_2) - ii(x_2, y_1)$. Please refer to Sec. 4.1.5.

(e.g. [550]) and the usage of contextual features such as head and shoulder locations (e.g. [298, 64]) are promising research directions for further improvements of appearance-based face detection approaches.

4.1.5. Boosting-Based Face Detection Approach of Viola and Jones [504]

The face detection approach developed by Viola and Jones [504, 503] is nowadays provably the most widely used and adapted one. They introduced the *integral image* representation of input images that enables very fast feature computation. The integral image contains at each position the accumulated sum of all pixel intensities with lower or equal index:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'), \tag{4.1}$$

where ii is the integral image and i the normal image (intensity values). The integral image can be computed in one pass over the input image. Using this representation, the energy (i.e. the sum of intensity values) of an arbitrary rectangle within the input image can be computed in constant time, as Fig. 4.1 demonstrates. This is used to compute a large number of simple features where the energies of neighboring rectangles are combined additively or subtractively; examples of these features are depicted in Fig. 4.2. This computation is very fast, as each feature can be computed in constant time at any position and scale.

Viola and Jones [504] used a variation of *AdaBoost* [167] to train a face/no face classifier based on a large training set of positive (face images) and negative (other images) examples. They consider a very large number of candidate features where each feature is used by a weak classifier that tries to distinguish between faces and non-faces based on a simple threshold on the feature value. The AdaBoost learning algorithm is utilized to built a strong classifier which is a combination of several of these weak classifiers. This also corresponds to a feature selection, because each weak classifier is based on exactly one feature.[1]

[1] Viola and Jones [504] reported experiments where 6,060 features out of 160,000 candidates were selected.

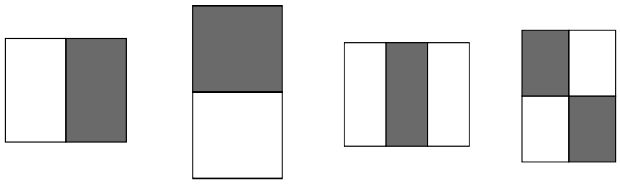


Figure 4.2.: Example illustration of the features Viola and Jones [504] used in their face detection approach. These features calculate the differences between the energy of the pixels under the dark rectangles and the energy of the pixels under the light rectangles. Please refer to Sec. 4.1.5.

To speed up the face detection, Viola and Jones [504] used an attentional cascade of classifiers. Each classifier is trained with AdaBoost and considers only those parts of the input image that were not rejected by the previous classifier. The classification thresholds of the classifiers are chosen such that each classifier has a very good true positive rate ($\approx 99.9\% - 100\%$ of real faces are detected) at the cost of a poor false positive rate (only $\approx 50\% - 80\%$ of non-faces are rejected). The classifiers are of increasing complexity in terms of the number of features they consider. Thus, the vast part of non-faces is rejected by the simple, but fast classifiers in early stages, whereas the more complex classifiers at later stages are only utilized to distinguish between real faces and other patterns that were not rejected so far. As usually the great majority of subwindows of the input image that are searched for faces do not show faces, this classification approach is very fast, because the more complex classifiers are activated very infrequently only.[2]

In recent years, numerous modifications and extensions of this face detection technique have been proposed. One line of extensions concerns the computed features. Lienhart and Maydt [328] added center-surround features and also 45° rotated variants of all features and achieved a 10% improvement of the false positive rate. Other feature variations were suggested by Li *et al.* [321], Jones and Viola [260], Mita *et al.* [361], and Meynet *et al.* [360] for instance. Improvements of the boosting learning algorithm are a second type of modifications that have been frequently presented by researchers. Wu *et al.* [523] and Mita *et al.* [361] suggested to use *Real AdaBoost* [442] instead of the original algorithm [167], Lienhart *et al.* [327] advocated *Gentle Adaboost* [171], and Li *et al.* [321, 320] introduced *FloatBoost*. Jang and Kim [251] optimized the classifier cascade with evolutionary algorithms, other researchers improved it by incorporating knowledge gained from classifications performed earlier in the cascade [523, 528]. Also the online adaptation of boosted classifiers for changed environments has been investigated [231]. Please refer to the survey of Zhang and Zhang [551] for a detailed discussion of these and other extensions of the face detection technique of Viola and Jones [504]. Furthermore, several researchers [64, 474, 512] improved the approach by incorporating skin-color detection as a post-processing step. We discuss one of these approaches in the next section.

[2] Viola and Jones [504] achieved state of the art face detection accuracy with a cascade of 38 classifiers, which performed significantly faster than previous approaches.

4.1.6. Encara Face Detection Approach of Castrillón *et al.* [64]

Castrillón *et al.* [64] developed the *Encara* face detection system for real-time face detection and tracking in video streams. It utilizes an extended version [327] of Viola and Jones' [504] boosting approach, implemented in the *OpenCV* library [40], for the initial detection of faces in two ways: one classifier cascade detects faces as usual [504, 327], while a second one is dedicated to the detection of heads and shoulders [298]. The latter is especially useful for low resolution images and non-frontal faces.

Once a (frontal) face has been detected, the locations of the eyes are searched using a multilevel eye detection method. The first step is to detect skin-colored blobs within the face region, accompanied by a heuristical elimination of non-face areas (e.g. neck) and a rotation into a vertical position by means of ellipse fitting [467]. Next, eye candidates are searched as relatively dark areas and also by Viola-Jones-based eye and eye pair detectors. The found eye coordinates are used to normalize the face bounding box to a standard size, before two final PCA-based verifications are performed: the appearance of both the eye region and the whole face are checked by projection into the PCA space and evaluation of the reconstruction error [220] and SVM classification, respectively.

As the face detection method of Castrillón *et al.* [64] focuses on video streams in particular, they apply several heuristics to track or quickly redetect a face that has been found in previous frames. These heuristics comprise a fast tracking of the eyes [200], an application of the two Viola-Jones-based detectors for faces and also heads and shoulders, restricted to the rough region of the previously detected face, a detection of skin-colored blobs [522], and a tracking of the recorded face pattern found in the previous frame [200]. All of these heuristics are executed in the given order until the face was successfully re-detected. Additionally, every five frames the whole image is searched for new faces as described above. Because the tracking and skin-color-based re-detection of faces are not as reliable as the Viola-Jones-based initial detection, those techniques are utilized only if a face with eyes was found in preceding frames. Thus, they help to keep track of a face in case one is already very confident that there actually is one present, without the comparatively high risk of false positive detections which their usage for initial face detection would accompany. Castrillón *et al.* [64] presented an experimental evaluation of their approach that demonstrated a better face detection rate and a significantly lower processing time on average, compared to the original approach of Viola and Jones [504].

4.2. Head Gesture Recognition

The common approach for head gesture recognition is to recover the head *pose* from input images and estimate the head *gesture* from the pose variations over time afterwards. Therefore, we focus on visual head pose estimation techniques, where Murphy-Chutorian and Trivedi [368] presented a comprehensive survey. They classified the existing state of the art approaches into eight categories, reviewed their theory and compared the achieved results. We follow their classification in our discussion here and consider appearance template methods, detector arrays, nonlinear regression methods, manifold embedding methods, flexible models, geometric methods, and tracking methods in the subsequent sections. Approaches of the eighth category, hybrid methods, combine several techniques form the first seven categories. We do not discuss them explicitly here, but refer to the respective section of the survey of Murphy-Chutorian and Trivedi [368].

4.2.1. Appearance Template Methods

Appearance template methods [368] use example images of faces under different head orientations as prototypes for a head pose classification. Typically, a test image is classified into the discrete head pose class of the best-matching retype. Beymer [30] used a similarity measure based on normalized cross-correlation for this classification, whereas Niyogi and Freeman [379] used the mean squared error in a sliding window over the image. To increase robustness, the images can be filtered to enhance prominent features such as horizontal or vertical lines, for instance by a convolution with Gabor wavelets as Sherrah *et al.* [456] did.

Besides its simplicity, an advantage of this approach is that the prototype images can be extended or exchanged on the fly, allowing for online adaptations to changing environments [368]. The most serious disadvantage is the difficulty to get a set of prototypes with good generalization properties, as differences due to the identity of a person can easily outweigh the differences induced by varying head poses [368]. One way to address this issue is to use a very large number of prototypes of different people to achieve a reasonable coverage of identity-induced variations. However, this leads to a high computational cost as the target image needs to be compared to every prototype. Niyogi and Freeman [379] tried to overcome this by using a tree-structured vector quantization technique where not every prototype is considered, but a fast heuristic is used to decide which subtree of prototypes is traversed at each stage of a recursive search during the classification. Ng and Gong [373] utilized a SVM for face localization and used the support vectors as appearance templates, which yields a significant reduction of the number of prototypes.

4.2.2. Detector Arrays

Detector arrays [368] are also appearance-based, but instead of using the face images directly as prototypes, they apply machine learning techniques to train several detectors to distinguish faces in a certain head pose from non-faces and also faces in another pose. One way to construct such a detector array is to train the detectors with the boosting approach of Viola and Jones [504] (please refer to Sec. 4.1.5), where the positive examples for each detector are only faces of a certain orientation. This was done by Zhang *et al.* [554], who used *FloatBoost* [320] instead of *AdaBoost* [167] and a Bayesian network to fuse the detection results of multiple cameras. Furthermore, they applied a HMM to recognize temporal changes in the detected head poses. Rowley *et al.* [424] developed an in a sense inverse approach where a "router" network first determines the orientation of an image patch, assuming it contains a face, and subsequently the image patch (rotated to normal upright position) is conveyed to one or more detector networks which decide whether the image actually shows a face.

A practical issue with detector arrays is the partitioning of the training data for the single detectors. Besides a large number of non-faces images, also face images of different orientations need to be included in the negative examples, as each detector is supposed to detect only faces of a certain orientation. This limits the number of detectors and thus discrete orientations that can be recognized, because the positive and negative face examples for two neighboring orientation classes are very similar, the more the finer the resolution in terms of number of orientations is, thus the training of high-quality detectors is difficult [368].

4.2.3. Nonlinear Regression Methods

Nonlinear regression methods [368] perform a continuous or discrete head pose estimation by learning a mapping function between the input data and the associated orientations. Several approaches used neural networks, in particular multilayer perceptrons (MLPs) [34], to realize this nonlinear mapping. Seemann *et al.* [452] utilized three-layer MLPs to estimate the pan, tilt, and roll angles of a head image. Voit *et al.* [505] used the same network type, but extended the method by a Bayesian filter framework to fuse the results gained from multiple cameras. Tian *et al.* [485] also integrated images of several low-resolution cameras to estimate the head pose. A different kind of neural network, a local linear map (LLM), was deployed by Rae and Ritter [412], based on a feature extraction with Gabor filters.

Another popular approach is support vector regression (SVR) [124]. Li *et al.* [322] used SVR and edge images to estimate the pose of an input image patch. Based on the estimated pose, a dedicated face detector based on eigenfaces [496] and SVM classification is applied to verify or reject a face in the input image patch (in principle similar to the approach of Rowley *et al.* [424] discussed above). Murphy-Chutorian *et al.* [367] combined SVR with localized gradient orientation histograms that lead to a dimensionality reduction (which is important for a successful SVR in many cases [368]). Other researchers performed a dimensionality reduction by PCA [322], Gabor wavelet networks [297], or by the concentration on a few salient facial features, for instance. For the latter, Gourier *et al.* [192] used locally normalized Gaussian receptive fields to find salient facial structures such as eyes, nose, mouth and chin. Ma *et al.* [341] presented a nonlinear regression method that uses a relevance vector machine [487] to learn the relations between the position of facial feature points and the 2D head pose.

Most regression methods of visual head pose estimation require a relatively precise localization of the head [368]. Osadchy *et al.* [383] tried to enhance the robustness by using convolutional networks which feature a reasonable local shift invariance, combined with multiscale images to achieve a certain size invariance, and performed the classification with energy-based models [315]. Recently, Fanelli *et al.* [155] presented a robust head pose estimation approach based on 3D depth information and random regression forests [42].

4.2.4. Manifold Embedding Methods

Manifold embedding methods [368] project the high-dimensional image data onto a low-dimensional manifold, assuming that the actual number of intrinsic dimensions of pose variations is comparatively low. These methods are related to the regression methods discussed above, as regression can be used to estimate the orientation from a point on the manifold; in fact, several methods can be classified into both groups (e.g. [383]). A key ingredient of these methods is a significant dimension reduction to construct the manifold of pose variations. McKenna and Gong [353] used PCA to do this, after a preprocessing with Gabor wavelets, and subsequently applied template matching for the classification. Sherrah *et al.* [456] demonstrated that PCA can be used to enhance pose differences and suppress undesired identity-induced similarities.

Nevertheless, variations due to different identities, lighting conditions, etc. are intermixed with variations caused by different poses. To cope with this problem, Srinivasan and Boyer [469] used pose-specific eigenspaces where a target image is projected onto all of them and classified into the discrete pose class with the highest projection energy. Each pose-specific

eigenspace is supposed to cover the variations due to other sources than pose. Li *et al.* [319] deployed a similar method, but utilized a SVM classifier for the final pose estimation.

However, according to the study of Wu and Trivedi [524], PCA is inferior to other methods for manifold embedding for head pose recognition, for instance (kernelized) linear discriminant analysis (LDA) [127]. In general, nonlinear methods such as Laplacian eigenmaps [28] and locally linear embedding [422] yield superior results. Li *et al.* [323] increased the robustness by a piece-wise linear approximation of the nonlinear pose manifold, whereas Balasubramanian *et al.* [16] presented a biased manifold embedding technique that effectively incorporates the pose information into the embedding process, which leads to a performance improvement.

4.2.5. Flexible Models

Flexible models [368] use a nonrigid head model that matches a set of feature points to the respective locations in the target image. This matching is learned from training data, where a set of training images with annotated feature points is required. Due to their deformable nature, flexible models can significantly improve the image registration compared to rigid models such as classical template matching (please see Sec. 4.2.1). Krüger *et al.* [296] used elastic graph matching techniques [301] for head pose estimation. For each pose, an elastic bunch graph [301] is compared to the target image via template matching. At the feature point locations, local feature descriptors based on "jets" of Gabor wavelets are computed. Wu and Trivedi [524] incorporated a similar approach in their head pose estimation framework. At first, the rough head pose is estimated using a subspace analysis based on Gabor wavelets. Afterwards, this pose estimate is refined by applying elastic bunch graph matching.

Another popular method are active appearance models (AAMs) [90], a derivative of active shape models (ASMs) [86]. These models learn the modes of shape and texture variation of faces from training images using PCA and apply an iterative search procedure to align the feature points to a target image. Lanitis *et al.* [314] used ASMs to recover the head pose from the principal components of shape variation that capture the largest variance. Baker *et al.* [15] used AAMs to track the head of car drivers. Cootes *et al.* [93] demonstrated that a wide range of head rotations can be captured by a small set of view-dependent AAMs. Gui and Zhang [201] also applied AAMs to get a 2D estimate of the head pose and subsequently utilized structure from motion techniques [455] to estimate the 3D pose. AAMs can yield a very accurate pose estimation. However, they require images of comparatively high resolution to perform the fitting. Furthermore, all feature points need to be visible, thus the range of allowed head rotations is limited. The last limitation has been addressed by extensions of this approach [199]. Please refer to Sec. 4.5.1 where we discuss AAMs in more detail.

4.2.6. Geometric Methods

Geometric Methods [368] exploit geometric relations of facial features to estimate the head pose. Gee and Cipolla's [186] head pose recognition method is based on the estimation of a facial normal using five points located at eyes, mouth, and nose. Horprasert *et al.* [226] used five different points and exploited projective invariants to recover the pose. Cordea *et al.* [94] fitted an ellipse to the face based on intensity gradients and color distributions inside and outside the ellipse. They tracked the head using a linear Kalman filter [264] and recovered the pose from the ellipse parameters. Wang and Sung [511] used geometric relations of eyes

and mouth corners to estimate the 3D pose of a head, whereby they utilized an EM-algorithm [112] to adapt to different persons and facial expressions. Geometric methods are typically very efficient in terms of computational cost, but require a very precise localization of the used facial feature points, which might be an issue for their practical applicability.

4.2.7. Tracking Methods

In contrast to the methods discussed so far, tracking methods [368] do not use single images to recognize the head pose, but estimate it from the relative motion between consecutive frames. The typical approach is to map observed 2D image features to a 3D head model, where the pose can be recovered from the parameters of this model resp. mapping. Yang and Zhang [536] used feature point correspondence techniques to estimate the 3D head pose with a stereo camera. Ohayon and Rivlin [381] estimated the camera pose in order to perform the mapping from the 2D feature points to the 3D head model and determined the pose by a backprojection of the model to the input image and error minimization. Haro *et al.* [209] applied a 3D-textured polygon model and performed the mapping by a gradient decent technique. Malciu and Preteux [342] realized a 3D pose estimation by a combination of motion and texture features, where the mapping is computed by an iterative optimization (a variant of the simplex algorithm). They improved the speed of their method by incorporating an interpolation based on optical flow.

Several other researchers used a particle filter [120] resp. condensation [241] framework for head pose detection. In these methods, a set of particles represent different assumptions about the object state (i.e. the head pose). Using a generative model, an image of the head can be generated for each particle, representing how the head might appear in case the pose associated with this particle is the true head pose. These images are compared to the input image to infer the pose and update the particle set [368]. Oka *et al.* [382] used particle filters equipped with an online control of the diffusion of the particle set to adapt to different velocities of head movements. Dornaika and Davoine [119] combined particle filters with an online appearance model and incorporated both observation and state transitions models to improve efficiency.

There are many other methods, for instance the approach of Zhao *et al.* [555] who introduced a 3D head tracking method that uses image registration based on SIFT features [338]. Tracking methods can produce very good results, but require a suitable initialization at the start or when the pose got lost.

4.3. Eye Gaze Recognition

Morimoto and Mimica [363] presented a review of several eye tracking approaches. Generally, the best performances are achieved by intrusive eye gaze trackers [363]. However, as these methods require the target persons to wear special contact lenses or to place electrodes around the eye, they are not applicable in the human-robot interaction scenarios we are interested in, thus we do not consider them here.

Another large group of method uses infrared light to enhance the intensity contrast between pupil and iris [363]. Yoo *et al.* [544], for instance, reported a real-time eye tracking system for people sitting in front of a computer monitor. Although such a system could be incorporated

in a robot, we are mainly interested in eye gaze recognition methods that use normal vision input, hence we do not discuss these approaches in detail.

Furthermore, in practice, chances in gaze directions are often accompanied be head movements. In this case, the approaches presented in the previous section can be utilized to get a rough estimation of the eye gaze (the gaze direction has to be aligned with the head pose anyway). A more precise gaze direction estimation usually requires images of a comparatively high resolution, such that geometric analyzes, etc. can be performed for the eye image. We briefly consider several approaches for this purpose below.

Wang and Sung [510] presented a system that uses the geometric relations of iris and eye corners, evaluated in a zoomed-in image of one eye, to robustly estimate the eye gaze. Nikolaidis and Pitas [378] developed a gaze recognition method where a Hough transform is combined with template matching and active contour fitting to detect facial features. To estimate the gaze direction, they exploited symmetry properties of the face. A stereo vision system for real-time head pose and eye gaze estimation by means of 3D eye corners and pupils tracking via template matching was described by Newman *et al.* [372]. Baluja and Pomerleau [17] investigated the classification of eye images with neural networks. Ishikawa *et al.* [242] used an AAM to locate the eye region and a subsequent ellipse fitting and template matching for gaze estimation, whereas Ivan [244] directly utilized AAMs to model the eye. Varchmin *et al.* [500] developed a system that combines eigeneye analysis, nose and mouth detection (for head pose estimation), and a series of neural networks to estimate the gaze direction of a user.

4.4. Facial Expression Recognition

A very large number of facial expression recognition approaches has been developed in the last decades. There are several different criteria that can be used to classify them. In their survey, Pantic and Rothkrantz [392] distinguished methods that operate on static images from methods operating on image sequences. In both categories, they further discriminated between template-based techniques and feature-based techniques. In this classification, template-based models use a holistic face representation that is either fitted to the target image or tracked in the input image sequence. Analogously, feature-based methods find or track certain facial features in the images. Regarding the used classification techniques, they distinguished template-based, neural-network-based, and rule-based approaches.

Fasel and Luettin [158] also presented a survey of automatic facial expression recognition techniques. They divided the utilized methods along three lines: deformation extraction vs. motion extraction, holistic representations vs. local representations, and model-based vs. feature-based approaches. Similarly to Pantic and Rothkrantz [392], they distinguished between spatial classification methods operating on single images and spatio-temporal methods considering temporal dynamics. In many cases, an approach could be classified into more than one category. In fact, some approaches were categorized differently by Pantic and Rothkrantz [392] on the one hand and Fasel and Luettin [158] on the other hand, in particular regarding the distinction of model- and feature-based methods. Therefore, we use a coarser categorizations here, separating only static approaches that operate on single images from dynamic approaches that inherently process video data.

4.4.1. Static Approaches

The static approaches considered in this section operate on single images. Several researchers applied AAMs to recognize facial expressions, these include Lanitis *et al.* [314], Edwards *et al.* [129], and more recently Lucey *et al.* [339] and Rabie *et al.* [410]. Huang and Huang [232] used a point distribution and gray-level model to derive 10 action parameters which were subsequently used for the expression classification. Hong *et al.* [224] classified facial expressions by evaluating the similarity of a target image to a set of person-specific facial expression models by means of elastic graph matching. Yoneyama *et al.* [543] used optical flow to compute the distortion parameters of an input image, compared to a neutral image as baseline model, and classified these parameters with hopfield networks.

Besides these holistic approaches, also local methods were proposed. Pantic and Rothkrantz [391] combined a frontal-view face model consisting of 30 features with a side-view model of 10 face profile points. They utilized several local feature detectors for the model fitting, for instance a neural-network-based eye detection and a rule-based-mouth geometry estimation. Kobayashi and Hara [290] used brightness distribution models for eyes, eyebrows, and mouth to detect these features by a cross-correlation technique. The classification was performed with a neural network. Lam and Yan [303] presented a method where 15 feature points corresponding to eye and mouth corners, eyebrows, nose, and other points on the face are located by dedicated local detectors. To perform the classification, they combined a 2D point-matching scheme with a correlation measure, which yielded considerably better results than the two methods individually. Ioannou *et al.* [240] presented a rule-based neurofuzzy network based on the model-based detection of inner facial features. Recently, Littlewort *et al.* [330] presented a facial expression recognition toolbox called *CERT* that performs local facial feature detection with dedicated boosting classifiers, feature extraction with Gabor filters, and a classification of action units (AUs) [139] with SVMs.

Fellenz *et al.* [160] investigated the classification of facial expressions using multilayer perceptrons (MLPs) [213] and a holistic preprocessing with Gabor filters. In the conducted experiments, they found this method to be superior to template matching and an alternative PCA-based feature representation, in particular regarding the generalization performance. Padgett and Cottrell [390] applied a PCA locally to regions around the eyes and mouth and used neural networks to classify facial expressions. Dailey and Cottrell [103] performed very similar experiments using an ensemble of nonlinear networks and found Gabor filters and local PCA representations to perform equally well. This contrasts the results of Bartlett [22] whose experiments demonstrated superior performance of Gabor representations and independent component analysis (ICA) [238], compared to local PCA and several other features.

4.4.2. Dynamic Approaches

Dynamic approaches do not operate on single images, but make use of the dynamic structure of videos. Sebe *et al.* [451] employed a piecewise Bézier volume deformation (PBVD) [479] to recognize local deformations of facial features. They applied and compared several classifiers, including generative Bayesian networks and decision trees. Terzopoulos and Waters [483] introduced an approach where a generative face model is combined with the tracking of facial features (eyebrows, nose, and mouth) using snakes. A multistage recognition model was developed by Yang *et al.* [535]. Based on a feature extraction with Haar-like features, they performed a clustering of temporal patterns and constructed dynamic binary patterns. The

classification was done with a boosting method. Donato *et al.* [116] presented an experimental comparison of several approaches for action unit classification in image sequences.

A large number of dynamic facial expression recognition methods is based on optical flow. Lien [326] combined a dense optical flow tracking with PCA and performed the classification with HMMs. Cohn *et al.* [84] developed a facial feature tracking method based on local optical flow around several feature points. Otsuka and Ohya [388] performed a facial expression classification with HMMs, based on motion estimation around eye and mouth by means of gradient-based optical flow and feature extraction with a 2D Fourier transformation. A combination of 3D deformable models and optical flow was presented by DeCarlo and Metaxa [108]. In their model, they utilized techniques based on anthropometry and edge forces to improve the alignment.

4.4.3. Descriptive Recognition and Interpretation

Fasel and Luettin [158] pointed out an important distinction regarding the outputs of facial expression recognition methods. Some approaches perform a *descriptive recognition*, mainly those which consider the recognition of action units (AUs) [139]. As discussed in Sec. 2.2.3, action units describe the visual appearance of a face and do not per se attribute a certain meaning to it. Among these approaches are the methods investigated by Bartlett *et al.* [21], who classified 20 AUs by means of Gabor filters and support vector machines (SVMs), Tian *et al.* [486], who presented a system that can recognize 16 AUs via geometric facial feature modeling and two neural network classifiers for the upper and lower part of the face, Lucey *et al.* [339], who utilized AAMs, and several others (e.g. [498, 339, 83]).

However, the majority of approaches does not perform such a *descriptive recognition*, but an *interpretation* of the displayed facial expressions. Many researchers investigated a classification into discrete categories, most often basic emotions [133]. Buenaposada *et al.* [53] built linear subspace deformation and illumination models and used a nearest-neighbor-based classifier for that purpose, whereas Lanitis *et al.* [314] used flexible models of shape and gray-level, also the related AAMs were utilized in other approaches [129, 410], as well as many other techniques, e.g. Haar-like features and dynamic binary patterns [535] and local facial feature deformations [451]. These interpretations of the facial expressions might be based on AUs. For instance, Prado *et al.* [406] recently performed basic emotion classification based on AU recognition in the upper and lower face with Bayesian networks and integrated the results with audio emotion recognition.

Some researchers considered other or additional categories. Sebe *et al.* [450] added interest, boredom, confusion, and frustration to the set of basic emotions. Kapoor and Picard [273] also investigated interest and boredom; Yeasin *et al.* [540] considered "levels of interest". El Kaliouby and Robinson [144] presented a system for the real-time inference of more complex mental states, namely agreeing, concentrating, disagreeing, interested, thinking and unsure.

Several recent approaches considered the recognition of facial expressions in terms of emotional dimensions. Caridakis *et al.* [62] and Fragopanagos and Taylor [166] investigated the recognition of valence and activation level with neural networks. Gunes and Pantic [202] used hidden Markov models (HMMs) and SVMs for the continuous prediction of five dimensions (arousal, expectation, intensity, power and valence). A classification approach based on fisher features and boosting for an activation-evaluation recognition was investigated by Zeng *et al.* [549]. Ioannou *et al.* [240] also performed an activation-evaluation recognition, but they used

a rule-based neurofuzzy network for this. McDuff *et al.* [352] investigated valence recognition using facial action unit spectrograms, where they evaluated the performance with several classifiers. Gunes and Pantic [202] considered the recognition of five dimensions (arousal, expectation, intensity, power, and valence) from head gestures. In total, however, the recognition of valence and other dimensions is not as intensively researched as the classification of discrete emotion categories.

4.4.4. Current Research Trends

When one compares the most recent approaches to facial expression recognition with somewhat older ones, it becomes evident that a paradigm shift is in process since a few years. While the typical approach used to be to consider posed facial expressions [548], the investigation of spontaneous, authentic ones continues to receive increasing research attention. For instance, Valstar *et al.* [498] showed that genuine and posed smiles can be distinguished automatically. Sebe *et al.* [451] investigated the classification of authentic basic emotions in a video kiosk scenario. Bartlett *et al.* [21] performed AU recognition on a database of subjects engaged in social or political discussions, Lucey *et al.* [339] also investigated the recognition of spontaneous AUs.

Zeng *et al.* [548] presented a comprehensive survey on this topic. The shift towards this investigation of spontaneous facial expressions is also accompanied by the need for other means of interpretation beside the basic emotions studied most often, as they are not well-suited for most interaction situations [548], which become a focus of investigations. In the conclusion of their survey, Pantic and Rothkrantz [392] stated that *all* of the considered approaches performed a classification into basic emotions. In the concluding remarks of their survey, Fasel and Luettin [158] noted that none of the surveyed approaches considered facial expressions during conversations. Fortunately, the survey of Zeng *et al.* [548] shows that meanwhile, several years later, this started to change.

4.5. Face Representation and Facial Feature Extraction

The discussions in the previous sections show that many face representations and feature extraction methods have been used for the recognition of head gestures, eye gaze, and facial expressions. Several of these methods turned out to be applicable for all three cases, making them well-suited for the investigations of FCSs in this work. The approaches discussed below are among those. In the following Sec. 4.5.1, we describe active appearance models (AAMs) in some detail, because it is the main feature extraction method we are going to use in the next two chapters. Furthermore, we also briefly discuss constrained local models (CLMs) and Gabor energy filter (GEFs) in Sec. 4.5.2 and 4.5.3, respectively, because we are going to conduct some additional experiments with these features, too.

4.5.1. Active Appearance Models

The *active appearance model* (AAM) [87, 90] is a generative model for images of deformable objects. Based on a set of training images with annotated feature points that define the shape of the shown face (please see Fig. 4.3), a shape model is constructed as follows: All training

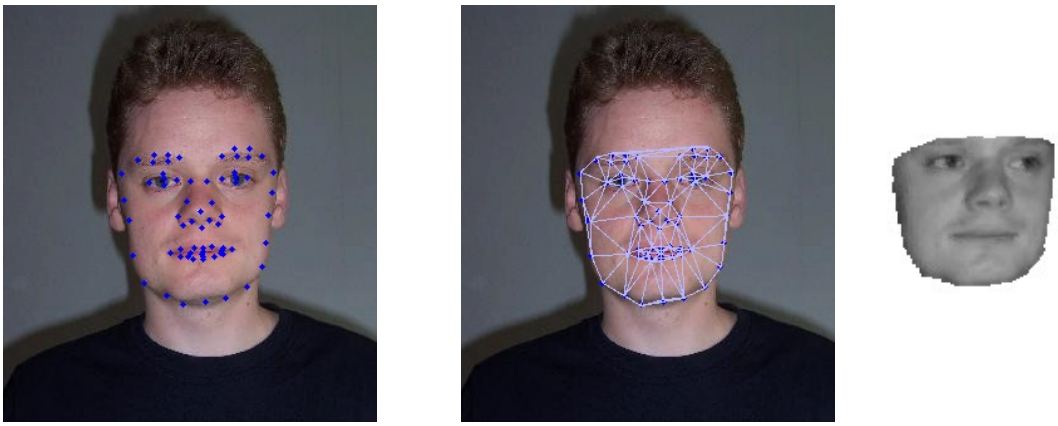


Figure 4.3.: Example images illustrating active appearance models (AAMs). Left: A face image with annotated feature points which describe its shape. Middle: An example of a Delaunay triangulation [106] that is used to compute the texture warping. Right: An example of a face image generated by an AAM. Please refer to Sec. 4.5.1.

images are aligned to a common reference frame such that shape variations due to global rotation, displacement, and scaling are removed. This is achieved by iteratively applying a *Procrustes analysis* [275, 193] to the whole training data set, until the mean shape of the aligned images does not change notably any more (which usually requires two iterations only [37]). Subsequently, a linear subspace model of the shape is built by a PCA:

$$\vec{x} = \bar{x} + P_s \cdot \vec{s} \tag{4.2}$$

With this shape model, a shape $\vec{x}$ (a set of 2D feature point coordinates) can be generated using the shape parameters $\vec{s}$, the mean shape $\bar{x}$, and the model matrix P_s which consists of a certain number of eigenvectors corresponding to the largest eigenvalues, whose number depends on the desired level of variance preservation. Having the shape model, the texture model is constructed next. In this context, "texture" refers to the pixel intensities under the area covered with feature points (i.e. the convex hull of the feature points). The first step is to warp all training images to the mean shape, using a triangulated mesh computed by a Delaunay triangulation [106] and affine transformations of the texture under the triangles (please see Fig. 4.3). Afterwards, the intensity values are normalized to reduce the influence of global lighting variations, before a linear subspace model of the texture is built using a PCA likewise to the shape model:

$$\vec{g} = \bar{g} + P_t \cdot \vec{t} \tag{4.3}$$

Thus, a texture $\vec{g}$ can be generated form the mean texture $\bar{g}$, the texture model matrix P_t, and the texture parameters $\vec{t}$. The resolution resp. size of the textures is a model parameter that is manually specified beforehand. The shape and texture model are combined to yield a linear model of appearance:

$$\vec{c} = \begin{pmatrix} W \cdot \vec{s} \\ \vec{t} \end{pmatrix} = \begin{pmatrix} W \cdot P_s^T \cdot (\vec{x} - \bar{x}) \\ P_t^T \cdot (\vec{g} - \bar{g}) \end{pmatrix}, \tag{4.4}$$

where W is weight matrix to account for the different units of shape an texture, i.e. to scale the former to a comparable range of variation. This is necessary for a third PCA that can optionally be applied, to further compact the representation, assuming that shape and texture may be correlated [87].[3] After this step, a combined parameter vector $\vec{c}$ can be gained using the appearance parameters $\vec{a}$:

$$\vec{c} = P_c \cdot \vec{a}. \tag{4.5}$$

The image corresponding to given shape and texture parameters can be generated by generating the shape-free image g first and warping it to the shape x afterwards (please see Fig. 4.3), where

$$\vec{x} = \bar{x} + P_s W P_{cs} \cdot \vec{a} \qquad \vec{g} = \bar{g} + P_t P_{ct} \cdot \vec{a} \qquad \text{with } P_c = \begin{pmatrix} P_{cs} \\ P_{ct} \end{pmatrix}. \tag{4.6}$$

A particular strength of an AAM is its ability to fit to new images, given a suitable initialization. This fitting is performed by an iterative search algorithm that minimizes the residual resp. reconstruction error $\delta\vec{g}$ between the image $\vec{g}_m$ generated by the model and the target image $\vec{g}_i$ under the current feature point positions (both represented as shape-free textures):

$$\delta\vec{g} = \vec{g}_i - \vec{g}_m \tag{4.7}$$

In each iteration, trial modifications of the appearance parameters $\vec{a}$ in a certain direction are performed, testing different step sizes and keeping those parameters that yielded the best reconstruction error. This method is iterated until the error is small enough or does not change any more. A key issue is the determination of the direction which should be used for these parameter modifications. This is given by the following linear relation:

$$\delta\vec{a} = R \cdot \delta\vec{g} \tag{4.8}$$

The matrix R is computed offline during the AAM training, which allows for a very fast fitting process. For all training images, the "true" appearance parameters are known. These are slightly modified in various directions (either systematically or randomly), and the resulting effect on the reconstruction error $\delta\vec{g}$ (caused by the changes in $\vec{g}_m$ due to the different appearance parameters) is observed. After a sufficient number of $(\delta\vec{a}, \delta\vec{g})$ pairs has been observed, R is computed using multivariate linear regression. Thus, AAMs perform the fitting to new images based on the assumption that there is a linear relation between the spatial pattern of the current reconstruction error and the direction into which the appearance parameter should be changed in order to improve the fitting, and that this relation can be learned from the training data in advance. However, in practice, this is an approximation that is valid only if the current feature point positions are already relatively close to their "true" positions, such a good initialization is required (please refer to Fig. 4.4). Besides the appearance parameters $\vec{a}$, also additional parameters describing global scale, size, and position variations are considered in this training.

[3] The matrix W can be chosen such that the global variance of shape and texture is equal after the scaling. Without such a scaling, the parameters of the model with the significantly smaller variance might be rejected almost entirely by the third PCA.

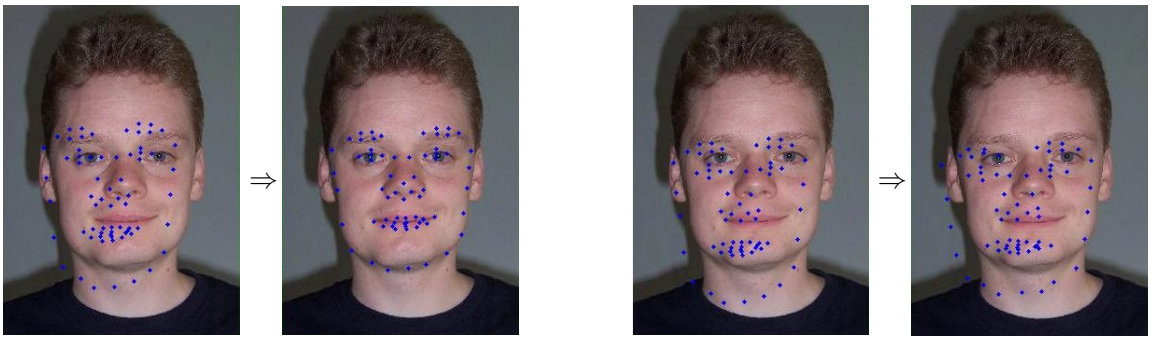


Figure 4.4.: AAM fitting examples. Left: This initialization is good enough, so the model can fit to the correct feature point positions. Right: This initialization is too far away from the correct positions, thus the fitting fails. Please refer to Sec. 4.5.1.

When combined with feature point displacements over time caused by head movements, these global transformation parameters allow an AAM to capture head gestures, provided that out-of-plane rotations are limited such that all feature points are still visible. In our object-teaching scenario, this is usually the case, as the subjects mainly directed their attention towards the robot. Due to the feature point spreading over the inner face area, an AAM can capture facial expressions very well. Although not explicitly modeled, an AAM can also account for eye gaze to some degree: pupil position and gaze direction are roughly captured by the position of and texture under the feature points that are dedicated to each eye. Moreover, changes in gaze direction are accompanied by head movements most of the time. Thus, AAMs appear to be well-suited to capture the kinds of FCSs that we investigate in the object-teaching scenario.

Cootes *et al.* [89] empirically compared the matching performance of AAMs and active shape models (ASMs). While ASMs yielded a more precise feature point localization, AAMs achieved a significantly better texture matching on face data. Edwards *et al.* [130] compared the fitting performance of nonlinear methods to linear ones and concluded that nonlinear techniques yield better results if the initialization is poor, but linear methods are superior if it is good, because they are less likely to drift away to a local minimum in this case.

AAMs have been used by many researchers and various extensions of the basic approach were proposed. Matthews and Baker [344] developed a new fitting algorithm based on inverse compositional image alignment and demonstrated its superior performance in terms of convergence properties and computational cost. Cootes and Taylor [88] also investigated an improved fitting method that tunes the AAM to the current image by updating the matrix R online, which can yield a better fitting accuracy at the cost of an increased runtime. Xiao *et al.* [527] presented an efficient combination of 2D AAMs and 3D models that can yield an improved fitting. Hu *et al.* [230] improved the fitting robustness by fitting an AAM simultaneously to images from multiple cameras. An AAM variant that can efficiently handle occlusion was developed by Gross *et al.* [199]. Doretto and Soatto [118] presented dynamic AAMs where in addition to shape and texture also motion between neighboring frames is represented. Bilinear AAMs were introduced by Gonzalez-Mora *et al.* [190] to support the decoupling of pose changes from facial expression and identity changes. Roberts *et al.* [419] incorporated a kernel method into the fitting procedure to reduce the sensitivity to outliers. A method to automatically select and place feature points by means of salient feature detection

Figure 4.5.: Example of a constrained local model fitted to a target image. Please refer to Sec. 4.5.2.

and tracking was investigated by Walker *et al.* [509].

4.5.2. Constrained Local Models

Several approaches that model shape variations by means of linear subspaces have been proposed, for instance active shape models [86] and constrained local models [513, 101]. Models of this class apply different kinds of local search strategies around feature points and use some global optimization method to jointly optimize the overall parameters, subject to certain constraints.

Saragih *et al.* [439, 440] showed that the optimization strategies of several of these models can be subsumed into a unified probabilistic framework they investigated. Furthermore, they presented a new, nonparametric optimization strategy based on isotropic Gaussian kernels. The conducted experiments demonstrated a very good performance of their method. Figure 4.5 shows an example for the fitting of this model to an input image.

4.5.3. Gabor Energy Filters

Gabor energy filters can be seen as a model for the complex cells in the visual cortex of the brain. Several researchers [301, 22, 116, 103, 518] have successfully utilized them for facial analysis tasks. The filters are computed from Gabor-based wavelets, where for each filter the real and imaginary parts of their responses resulting from the convolutions with the input data are squared and added in order to compute the energy. In experimental evaluations, a number of 40 filters at eight equally spaced orientations (at 22.5° intervals), each combined with five spatial frequencies with wavelengths of 1.17, 1.65, 2.33, 3.30, and 4.67 standard iris diameters,[4] were found to be well-suited for face recognition tasks.

[4] An iris diameter is the seventh part of the distance between the eye centers.

5. A Static Baseline Approach

Always try the simple things first.

— proverb

In this chapter, we investigate a simple static baseline approach for the automatic recognition of FCSs in terms of valence. The term "static" refers to the fact that this simple approach does not consider any temporal dynamics of the videos, but operates on frame level and treats each frame independently of all others. The main purpose of this investigation is to get first automatic valence recognition results in the object-teaching scenario, which shall serve as a baseline for the more sophisticated dynamic approach that is presented in the next chapter. The study of such a static approach is also motivated by psychologic research, where Bruce *et al.* [46] argue that humans also use a static representation of facial displays, in addition to a dynamic one.

The first Sec. 5.1 describes the utilized automatic face detection with Viola-Jones-based [504] techniques. The performed feature extraction by means of active appearance models, Gabor energy filters, and raw images directly is explained in Sec. 5.2. After a brief description of the deployed software in Sec. 5.3, the person-specific classification with support vector machines is evaluated in Sec. 5.4. Finally, Sec. 5.5 summarizes and concludes this chapter.

5.1. Face Detection

We compared the performance of three software libraries for face detection by visual inspection of the detection results when applied to the object-teaching videos:

- An implementation of Viola and Jones' [504] boosting approach with various modifications and extensions, with a particular focus on fast classifier training. This software was developed at Bielefeld University by Peters [400] as part of his diploma thesis. (Please refer to Sec. 4.1.5 and [400].)
- The *OpenCV* [40] implementation of Viola and Jones' [504] boosting approach (including the extensions of Lienhart *et al.* [327]). (Please refer to Sec. 4.1.5.)
- The *Encara* face detection software developed by Castrillón *et al.* [64] in an extended version which—in addition to the eyes—also locates mouth and nose. (Please refer to Sec. 4.1.6.)

We applied all three implementations to all scene videos in our object-teaching database and visually inspected the results, Fig. 5.1 shows typical example images of the detection results. This evaluation revealed that on average, the *Encara* software yielded the best face detection results at a sufficient speed, while the implementation of Peters [400] is the fasted approach. An additional advantage of *Encara* compared to the other two implementations is that it also detects eyes, mouth and nose. These inner facial features are used to improve the initialization

of the feature extraction (please refer to Sec. 5.2.1). Thus, we used the *Encara* software in our investigations of automatic FCS recognition presented in this and the following chapter.

Recently, Degtyarev and Seredin [110] empirically compared the performance of seven state of the art face detection software libraries on nine datasets. The commercial face detection software *VeriLook* [371] performed best in their tests, followed by the open source implementation of Viola and Jones' [504] boosting approach (including the extensions of Lienhart *et al.* [327]) in the *OpenCV* library [40]. The *Encara* software [64] performed even better than *OpenCV*'s boosting implementation on our object-teaching database, which demonstrates its high quality. Nevertheless, we found it beneficial to apply some additional postprocessing steps, as explained in the next section.

5.1.1. Postprocessing

In order to remove false positives, we rejected all detected faces of atypical dimensions, i.e. all faces smaller than 30×40 pixel or larger than 120×150 pixel. Face sizes out of this range were accepted only if a face of plausible size was found at a close location in 5 preceding frames. In this case the face was regarded as being real because a face is expected at this position.

Furthermore, we applied a postprocessing step that makes use of knowledge about our interaction scenario. In all considered interaction scenes, only one person is interacting with the robot, and this person roughly faces the robot almost all the time. Thus, we reanalyzed all detected faces of an interaction scene and identified a "main line" of faces, i.e. a (preferably long) sequence of plausible face detections over time with only slight changes in position and size from one frame to the next. This sequence is regarded as representing the real positions of the face of the robot's interaction partner, while all other faces that appeared outside of it are rejected. In the investigated object-teaching interaction scenario, all of these rejected faces were false positives as the robot's interaction partner was the only person close to the robot. However, even if they were real faces of people in the background, their rejection would not harm as the robot is only interested in its current interaction partner in the current scenario. In case of a scenario extension where multiple people shall interact with the robot at the same time this postprocessing needs to be refined, of course. One obvious way would be to extend it to identify multiple sequences of plausible face detections in parallel.

Besides the postprocessing step explained above which exploits the fact that the robot is interested in one person only, we also made use of the fact that this person faces the robot almost all the time during the considered interaction scenes. Thus, if the face disappears for just a few frames and reappears at roughly the same position, it is very likely that it actually was there all the time, but the face detector failed to locate it. We added these supposedly missing detections by linearly interpolating between the last detection before this gap and the first detection after it.

Most of these postprocessing steps require the modification of face detection results of previous frames based on the detection results for more recent frames. In the evaluations presented later in this and the next chapter, this was easily accomplished by revising the whole detection history after the complete interaction scene have been processed. Of course, this requires a waiting for and recognition of the end point of such an interaction. During the online performance of the robot without presegmented interaction scenes, this can be achieved by using a delay of a certain number of frames: within this time window, previous detection results may be modified before they are conveyed to the feature extraction and all subsequent

Figure 5.1.: Typical face detection results for object-teaching videos of the boost detector of Peters [400] (left), *OpenCV*'s boost detector [40] (middle), and the *Encara* detector [64] (right). Top row: an easy case where all three detectors find the face and *Encara* additionally locates eyes, mouth, and nose. Middle row: a more difficult case with a rotated face which is slightly out of the training range of Peters' boost detector, whereas the *OpenCV* detector finds the face but also an additional false positive, and the *Encara* detector locates a little too small bounding box without nose and mouth and rough eye positions only. Bottom row: a hard case where the head and shoulder detector of the *Encara* approach is able to detect the face roughly nevertheless, while the others cannot find it due to the occlusion caused by the hair. Please refer to Sec. 5.1.

processing steps. This would inevitably lead to a certain delay of the reaction of the robot to the displayed FCSs. However, this is not necessarily a negative point regarding the interaction, as FCSs usually require a certain time to unfold anyway, thus a too early interpretation might be harmful instead (please refer to the human recognition performance for different temporal contexts discussed in Sec. 3.5.2).

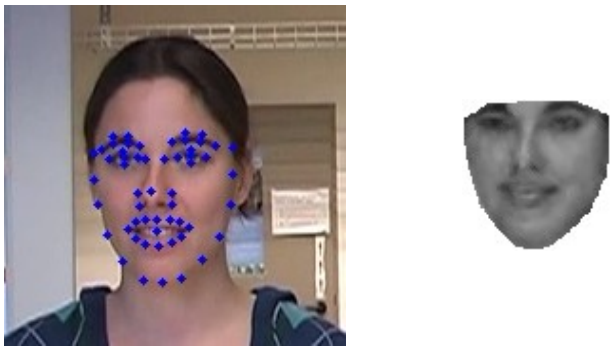


Figure 5.2.: Left: an AAM with 55 feature points fitted to a target image. Right: the image generated by this AAM. Please refer to Sec. 5.2.1.

5.2. Facial Feature Extraction

To gain the feature vectors that represent the single frames of the considered interaction videos, we used three types of feature extraction models in our investigations: active appearance models, Gabor energy filters, and raw face images. These feature extraction techniques are briefly discussed in the next sections, before their suitability for FCS recognition is empirically evaluated in Sec. 5.4.

5.2.1. Active Appearance Models

Our first choice for the extraction of facial features are active appearance models (AAMs) [90] as discussed in Sec. 4.5.1. There are several motivations for this choice:

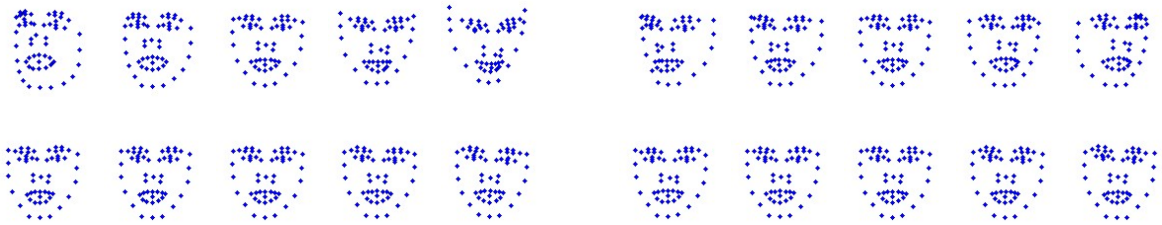
- Adult humans process faces in a *holistic* way, in contrast to a *parts-based* processing of most other objects (please refer to Sec. 2.1.2). Thus, a holistic face representation, like AAMs offer, appears to be appropriate for FCS recognition.
- Both shape and texture of faces seem to be important for human face recognition [461] (Please refer to Sec. 2.1.3). Thus, considering both in the feature extraction model appears suitable for FCS recognition.
- A recent study of Abiantun *et al.* [1] suggests that both shape and texture information are useful for automatic FCS recognition as well.
- AAMs can capture facial expressions and also eye gaze and moderate head movements to a reasonable degree (please refer to Sec. 4.5.1).
- As discussed in Sec. 4.2 to Sec. 4.5.1, AAMs have been successfully used for head gesture, eye gaze, and facial expression recognition by several researchers indeed.

For each person of the object-teaching study, we built an individual AAM from approximately 200 face images of this person with 55 hand-annotated feature points (please see Fig. 5.2). These images were manually selected in order to capture the variance in the facial displays as good as possible. Figure 5.3 depicts example illustrations of the resulting AAMs. The parameter vector of the model (when fitted to an image of a target video sequence) is used as feature vector for the respective frame. We used individual AAMs instead of generic ones because they usually yield considerably better fitting results [198].

The AAM fitting is initialized by an extended version of the initialization scheme developed by Rabie *et al.* [409]. The mean shape of the AAM is placed inside the detected face bounding

Variation of the appearance parameters (combined shape and texture parameters)

Variation of the texture parameters



Variation of the shape parameters

Figure 5.3.: Example illustration of the AAM for subject 01 of the object-teaching study. The upper block shows the model variance represented by the combined appearance parameters of shape and texture, the middle block the variance represented by the texture parameters, and the lower block the variance represented by the shape parameters. In each block, the first four parameters are varied and the effect is shown in a series of five images in each case: $\bar{x} - 2\sigma$, $\bar{x} - \sigma$, $\bar{x}$, $\bar{x} + \sigma$, and $\bar{x} + 2\sigma$ (from left to right), where $\bar{x}$ is the mean value of the respective parameter and σ its standard deviation. The top left series depicts the variantion of the first parameter, the top right the second one's, the bottom left the third one's, and the bottom right the fourth one's. The parameters are sorted according to the variance of the model training data they represent. It can be seen that the first two shape parameters mainly represent head poses, while the following ones correspond to more subtle variations in the face appearance. The shown texture parameters represent slight changes in illumination, but also to some degree eye gaze variations and mouth movements. Thus, the first four combined appearance parameters of shape and texture represent head movements, eye gaze, and mouth movements as dominating facial expressions, which is in accordance with the analysis of the FCS displays carried out in Sec. 3.4. Please refer to Sec. 5.2.1.

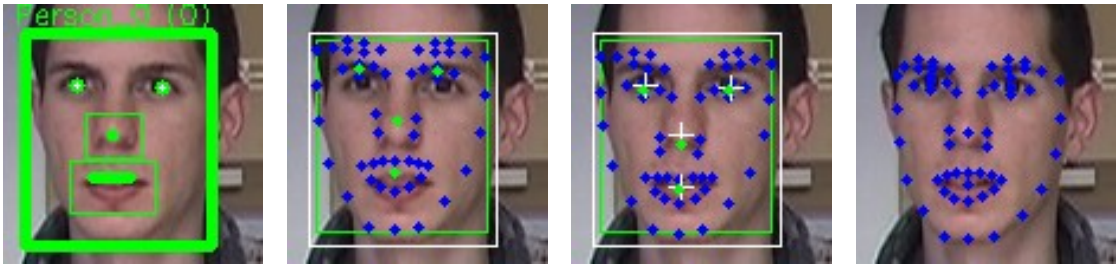


Figure 5.4.: Example illustration of the initialization method developed by Rabie *et al.* [409]. First image from left: This method uses the detected face bounding box and the locations of eyes, mouth, and nose, provided by the *Encara* face detector [64], for the initialization of the AAM fitting. Second image: The mean shape of the AAM is placed inside the face bounding box (white box), with a certain (predefined) size adaptation (green box). Third image: The feature points shown in green correspond to the positions of eyes, mouth, and nose. They are displaced to roughly match the locations of this features provided by the face detector (shown as white crosses). All other feature points (shown in blue) are also displaced, depending on their distance to those green feature points, weighted by a Gaussian. This yields an improved initialization. Fourth image: This improved initialization leads to an accurate fitting of the AAM to the face. Please refer to Sec. 5.2.1.

box, which is a reasonable initialization for the AAM in itself. This initialization is improved by a Gaussian distortion of the feature points. Based on their distances to eyes, nose, and mouth (whose positions are provided by the *Encara* face detector) the feature points are moved: those feature points that correspond to the centers of eyes, nose, and mouth are placed roughly at the detected locations of these features, and the surrounding feature points are also moved towards these locations by means of a Gaussian warping where the amount of their shift depends on their distance to eyes, nose, and mouth. Figure 5.4 shows an example of this initialization. In the experiments of Rabie *et al.* [409], this initialization improved a subsequent classification of facial expressions and also the identification of faces and performed better than a few other, similar initialization methods that were tested. Please refer to their paper [409] for further details.

In case the fitting using this initialization is good enough, i.e. the reconstruction error of the generated image is below a predefined threshold, it is accepted and used to get the feature vector. Otherwise, a series of slight global scale and rotation transformations is applied to the whole feature point set to get additional initializations, which are used until a fitting is good enough, or the whole series has been processed (in that case the fitting with the lowest reconstruction error is used). Before the initialization method described above is applied, the result of the AAM fitting from the previous frame (if available) is used as initialization for the current frame, the initialization method is only needed if this fails.

5.2.2. Gabor Energy Filters

We also applied a bank of 40 Gabor energy filters (GEFs) as second feature extraction method (please refer to Sec. 4.5.3. This bank contains filters at eight equally spaced orientations (at 22.5° intervals), each combined with five spatial frequencies with wavelengths of 1.17, 1.65,

2.33, 3.30, and 4.67 standard iris diameters.[1] This filter design was also used by Whitehill *et al.* [518] and found to be well suited for face recognition [116, 518]. The response images of the filters, when convolved with the face image (i.e. the image under the detected face bounding box), are downscaled and concatenated to form the feature vector of the respective frame.

5.2.3. Raw Face Images

For performance comparison, we also used the face images directly as features. The image under the detected face bounding box was downscaled and the RGB color values were concatenated and used as feature vector for the respective frame. Additionally, the image was also converted to gray scale and the resulting intensity values formed the feature vector.

5.3. Deployed Software

As already stated above, the *Encara* framework [64], written in C++, was used for the face detection. Additionally, the *OpenCV* [520, 40] implementation and Peters implementation [400] of Viola and Jones' [504] boosting approach were tested. The postprocessing of the detected faces was done by the *PseudoTrackingLibray*, a C++ software we developed for that purpose. The *BAM* software performed the feature extraction with AAMs. This C++ software was developed in previous work [304] and has been extended and modified for the experiments reported below. It uses the AAM implementation of the *Recognition And Vision Library (RAVL)* [66]. The GEF features [518] were implemented in Matlab, likewise to the raw image features. In both cases, the communication between the C++ and Matlab components was realized by the *matlabmm* and *matlab4iw* packages that have been developed in our group.

For the SVM classification, the *caiwicat* framework was utilized. Similarly to the *BAM* software, *caiwicat* is written in C++ and was originally developed in previous work [304] and has been revised and extended in this work. It uses the *Support Vector Machine Template Library (LIBSVMTL)* [395] as back-end implementation, which in turn is based on *LIBSVM* [70, 71]. Most of the aforementioned software was used in form of plugins for the iceWing framework [337, 336] which we utilized as general processing environment.

5.4. Evaluation

This section reports an investigation of several person-specific classifications of the *success* and *failure* scenes of the 11 people in the object-teaching database (please refer to Sec. 3.2). We used a support vector machine (SVM) [445] with radial basis function (RBF) kernel for the classification. The evaluation for all videos of a subject was conducted in a leave-one-out cross validation scheme: all frames of all videos except one were used for the training, then the excluded video was used as test data for the classifier. Section 5.4.1 investigates a classification by means of a simple majority voting over the frames of a video. Subsequently, this method is further simplified by representing each video by the mean vector of its frames only (Sec. 5.4.2).

[1] An iris diameter is the seventh part of the distance between the eye centers. In the experiments, the mean distance of the eyes detected by the *Encara* face detector was used for each person.

features	para.	grid values
AAMs	σ	0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1
	C	1, 2, 5, 10, 20, 50, 100
GEFs	σ	1, 2, 5, 10, 20, 50, 100
	C	1, 2, 5, 10, 20, 50, 100
Images	σ	0.1, 0.2, 0.5, 1, 2, 5, 10
	C	10, 20, 50, 100, 200, 500, 1000

Table 5.1.: Values of the SVM classifier parameters σ and C that were used in the grid searches for parameter optimization, each for different features: active appearance models (*AAMs*), Gabor energy filters (*GEFs*), and raw images (*Images*). The parameter ranges were empirically determined by preliminary tests. Please refer to Sec. 5.4.1.

Finally, the achieved results are compared to the human recognition performance in Sec. 5.4.3. Please refer to App. A.7 for a list of the feature vector dimensionalities in the classification experiments evaluated below.

5.4.1. Majority Voting Over Frames

We used a SVM to classify each frame of a test video independently of all other frames. The final classification result for the test video is formed by a simple majority voting over the classification results of the single frames; ties were counted as wrong classification. In a first experiment, we evaluate the performance of different feature variants. Subsequently, we investigate the variants that yielded good results in more detail.

Selection of Feature Variants

A 10-fold cross validation over all frames of all videos of a person was used to optimize the parameters of the SVM classifier, namely the RBF parameter σ and the regularization cost C. Table 5.1 lists the parameter values that were evaluated during the grid search. After the best parameters were found, a SVM was trained with these parameters using all frames of all training videos.

The classification results are shown in Tab. 5.2. Different variants of the features were investigated. We used AAMs that preserved 95% and 99% of the variance of their training data (rows *AAM-96* and *AAM-99* in Tab. 5.2, respectively).[2] Three variations of the GEF features were evaluated: the response images were downscaled to 4×4, 8×8, and 12×12 pixels; the corresponding results are shown in the rows *GEF-4/-8/-12* in Tab. 5.2. The direct usage of the image as feature vector was investigated in six variants where the image was scaled down to 8×8, 16×16, and 25×25 pixel, each for both gray level and RGB images (rows *RGB-8/-16/-25* and *gray-/8/-16/-25* in Tab. 5.2).

It is important to note that the purpose of these classification results is to compare the *relative* performance of the different feature variants when well-suited classifier parameters are used. Thus, a poor performance might not be caused by an inappropriate choice of parameters, but is rather intrinsic to the features resp. feature-classifier combination. Based on this

[2] All AAMs used a texture image size of 80×80 pixel.

features	all scenes		*success* scenes		*failure* scenes	
	mean	SD	mean	SD	mean	SD
AAM-95	63.6	23.1	54.8	27.1	69.8	23.9
AAM-99	**76.1**	10.3	66.6	18.0	83.2	12.2
GEF-4	**72.8**	12.3	65.7	28.4	76.3	15.5
GEF-8	73.1	11.6	66.5	27.9	76.0	15.4
GEF-12	71.3	12.9	64.4	27.9	74.9	17.3
RGB-8	72.5	13.9	63.8	28.1	77.6	14.6
RGB-16	72.1	14.3	65.9	28.0	74.2	17.8
RGB-25	68.2	16.7	60.9	29.7	70.8	27.1
gray-8	73.3	13.1	69.3	23.7	73.3	19.6
gray-16	**75.1**	12.5	67.5	25.3	79.0	14.6
gray-25	74.8	14.1	66.5	30.4	78.9	16.1
AAM-I	70.5	11.1	64.0	21.2	73.3	18.3

Table 5.2.: Mean value and standard deviation of the classification accuracies for all scenes, only *success*, and only *failure* scenes (distribution over the persons in the object-teaching database), in each case for different features. Please refer to Sec. 5.4.1 for an explanation of the listed feature variants and further details. The feature variants marked in bold are further investigated in subsequent experiments.

relative performance, the feature variants that were investigated in greater detail were chosen. However, these classification accuracies are not so much meaningful regarding their *absolute* value: as the parameter optimization was done using *all* videos of the respective person, it also included the particular test data for the single leave-one-out classifications. Despite the fact that the training itself was performed on the training videos only (without consideration of the test video), the test video had been "seen" by the parameter optimization before, thus the parameters might be tuned towards the test data.

The reason why we did it this way is the immense saving of training time: the time-consuming grid search for parameter optimization was necessary only once per person. In contrast, in a proper evaluation of the classification accuracies—as we report below—such a grid search is required before every single classification, considering the respective test data only. In case of the object-teaching database, this results in 447 grid searches (one for every *success* or *failure* scene in the database), compared to 11 grid searches in the method described above (one for every subject). This simplification is possible in this case, because we are interested in the *relative* performance of different features only, whereby for each feature type the same classifier and the same parameter optimization procedure is used. (Thus, the comparison is fair as *all* feature types benefit from knowing the test data during parameter optimization.) However, this method would not be valid if the actual performance of a classifier (in terms of *absolute* classification accuracies) was to be evaluated.

From Tab. 5.2 it can be seen that the achieved classification accuracies are rather low for a two-class problem. However, the classification problem is expected to be very hard, as the average human recognition accuracy is only 82% (please refer to Sec. 3.5.2). The best performance was achieved by an AAM with 99% variance preservation. But the raw image features compared surprisingly well to the AAMs. The reason for this is most likely that about 19% of the frames needed to be rejected from the AAM classification, because the

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	85	69	83	89	84	54	64	67	66	75	80	74.2	11.0
AAM – success	80	65	82	80	75	53	48	72	17	58	63	63.0	19.1
– failure	89	75	83	100	94	54	77	62	91	92	91	82.5	14.3
– all	72	66	80	95	88	61	66	78	54	58	75	72.1	12.7
GEF – success	80	53	86	95	88	80	44	84	0	50	58	65.3	27.9
– failure	65	83	72	94	88	38	84	69	83	67	86	75.4	15.6
– all	94	66	80	97	84	64	59	72	60	71	73	74.5	12.9
gray – success	93	76	86	95	75	73	32	81	17	58	54	67.3	24.8
– failure	94	50	72	100	94	54	81	62	83	83	86	78.1	16.6

Table 5.3.: Classification accuracies for different features where the classification was performend in a leave-one-out cross validation manner with parameter optimization on the training data prior to each training. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 5.4.1.

model fitting was too poor, mainly due to too large head rotations. When the performance of the raw image features is evaluated only on those frames that are actually used in the AAM tests, the classification accuracy decreases notably, as listed in row *AAM-I* in Tab. 5.2. This shows that also frames with a large head rotation contributed to the classification.

The GEFs yielded the worst classification performance. Theoretically, they are expected to outperform the raw image features. We surmise that compared to the amount of available training data, the dimension of the feature vectors is too high, even though the Gabor responses are highly downscaled (which might be a problem in its own), making it difficult to find appropriate class borders. (Please refer to App. A.7 for a list of the feature vector dimensionalities.) Following these considerations, it might be beneficial to use less filters with higher resolution in future work.

For the subsequent investigations, we consider the best performing variant of each feature only (marked in bold in Tab. 5.2), except for the GEFs, where we used variant *GEF-4* instead of *GEF-8* because of the lower feature vector dimensionality (640 compared to 2,560) and the only marginal difference in classification accuracy (0.3% means just one more video classified correctly).[3]

Feature Performance Evaluation

The second experiment investigates the feature variants chosen above in greater detail. As we are interested in the *absolute* classification performance this time, we conducted the parameter optimization prior to each training, using only the respective training set of videos for the grid search. The resulting classification accuracies are shown in Tab. 5.3. They are slightly lower than the ones in the feature selection experiment reported above, due to the different parameter optimization. The AAMs and the raw image features achieved similar results on average (74.2% and 74.5%, thus the raw image features classified one more scene correctly), whereas the GEFs were approximately two percentage points behind. However,

[3]However, the differences in the classification performance of the different features (the best variant in each case) are not statistically significant ($p > 0.4$ in all cases for both a two-tailed t-test and a Wilcoxon rank sum test).

	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**
classification accuracy difference	-9	3	3	-8	0	-10	5	-5	6	4	7
AAM: percentage of accepted frames	71	98	76	89	57	41	89	75	93	87	98

Table 5.4.: First row: the differences in the classification accuracies of the AAM features and the raw images features, for the 11 subjects of the object-teaching study (please see rows *AAM* and *gray* in the upper block of Tab. 5.3). Second row: the amount (percentage) of frames of the videos from the respective person that were accepted in the AAM-fitting and thus used for classification. There is a significant correlation between these quantities (Spearman correlation, $\rho \approx 0.73$, $p < 0.011$). Please refer to Sec. 5.4.1.

these differences are not significant.[4] In all cases the variance between the subjects was very high, comparable to the high variance of the human recognition performance, in part even higher (please see Sec. 3.5.2).

Comparing the classification performance of AAMs and raw images for the individual subjects, there is a significant correlation between the differences in classification accuracy and the amount of accepted frames in the AAM-fitting (Spearman correlation, $\rho \approx 0.73$, $p < 0.011$), please see Tab. 5.4. Again, this suggests that the missing frames that were rejected due to too poor AAM-fitting are the reason why the AAMs did not outperform the raw image features. Beyond that, we did not find a clear relation between differences in the AAM and raw feature performance and the kind of FCSs that are dominant for a person, as investigated in Sec. 3.4. As certain head poses are a main cause for AAM-fitting failures, a reasonable conjecture would be that people who use head gestures very frequently would come along with a relative high number of rejected frames und thus a relative poor classification accuracy, compared to the raw images features. However, this is not the case, rather the utilized AAMs can match the (frequently occuring) head gestures of some subjects very well while they fail to do so for others. Despite the relation between the amount of rejected frames and differences in the classification accuracies of AAMs and raw images, overall all features performed similarly in the sense that there are significant correlations between the features regarding the achieved results for the individual subjects (Spearman correlation, $\rho \approx 0.71$, $p < 0.02$ for *AAM* and *GEF*, $\rho \approx 0.95$, $p < 0.001$ for *AAM* and *gray*, and $\rho \approx 0.80$, $p < 0.01$ for *GEF* and *gray*).

In the following investigations, we focus on the AAM and raw image features only, because they performed better than the GEF features. Although only about eight percentage points behind the human performance, the achieved accuracies are rather low for a classification problem with two classes. We think that a main reason for this difficulty is the high intraclass variance, compared to the interclass variance. As a rough, but illustrative estimate of these variances, we computed the mean pairwise euclidean distances between all *success* and all *failure* frames separately (mean intraclass distance), and also the mean pairwise euclidean distance between all *success* and all *failure* frames (mean interclass distance) of a subject. These distances are listed in Tab. 5.5. The mean intra- and interclass distances are of comparable sizes, which indicates the difficulty of the classification problem. There is a highly significant correlation between the classification accuracies and the ratio of interclass to intraclass distance, the latter represented as the sum of the intraclass distances of the two classes (Spearman correlation, $\rho \approx 0.94$, $p < 0.001$ for AAMs, and $\rho \approx 0.98$, $p < 0.001$ for raw images). This supports the

[4] $p > 0.5$ in all cases for both a two-tailed t-test and a Wilcoxon rank sum test

features	01	02	03	04	05	06	07	08	09	10	11
– inter	26.8	21.3	24.5	29.9	37.5	46.8	29.0	27.4	23.0	29.4	24.6
AAM – success	25.5	23.0	26.9	30.9	39.9	44.3	27.4	29.8	22.1	23.0	25.4
– failure	22.1	17.6	19.3	21.8	28.7	47.2	29.3	20.3	21.8	33.0	17.1
– inter	2.73	1.80	2.43	2.92	2.95	3.23	2.19	2.43	1.48	2.75	2.25
gray – success	3.03	1.89	2.52	3.11	2.85	3.14	2.17	2.68	1.43	2.42	2.39
– failure	2.09	1.54	2.16	2.08	2.83	3.24	2.15	1.85	1.44	2.89	1.70

Table 5.5.: Mean intra- and interclass distances of the feature vectors, each for different persons and for AAM and raw image features. The *inter* rows show the mean interclass distances, i.e. the mean pairwise euclidean distances between all *success* and all *failure* frames. The *success* rows list the mean pairwise euclidean distances between all *success* frames, likewise do the rows *failure* for all *failure* frames; these are the mean intraclass distances of the respective class. Please refer to Sec. 5.4.1.

features	01	02	03	04	05	06	07	08	09	10	11
AAM – $\sigma \cdot 10^3$	5.52	7.59	2.17	3.56	0.98	0.77	5.09	9.48	3.26	2.91	11.78
– C	32.0	20.0	44.3	15.1	39.1	35.2	19.9	19.0	37.9	14.0	17.1
GEF – σ	8.97	11.21	6.96	8.78	5.63	9.11	9.82	9.48	8.63	5.21	13.22
– C	21.8	26.6	14.3	27.3	21.3	6.5	8.3	9.3	14.0	13.3	69.2
gray – $\sigma \cdot 10$	2.03	5.17	5.00	1.73	3.69	4.36	5.89	5.00	14.00	1.92	8.64
– C	47.6	14.8	10.7	52.2	12.2	13.6	12.0	10.7	15.4	22.9	29.0

Table 5.6.: Mean classifier parameters σ and C that were used for the training of all scenes of a person, for AAM, GEF, and raw image features. Please refer to Sec. 5.4.1.

hypothesis that a low interclass to intraclass variance ratio is a main reason for the frequent misclassifications in the investigated scenario.

Parameter Stability

Naturally, the leave-one-out cross validation tests yielded one parameter set (σ, C) for each scene, found via the grid search for parameter optimization on the respective training data. For the practical use of a classification system, usually a certain stability of the parameters is required, because a classifier trained with a particular parameter set is expected to give reasonable results on various test data. In order to test this stability, we performed another classification experiment where we used the mean σ and C values found during those grid searches for the training and classification of all scenes of a person. These mean parameters are listed in Tab. 5.6, the resulting classification accuracies are shown in Tab. 5.7. The classification accuracies are comparable to those achieved in the cross validation tests, for the most part even slightly higher. As the same parameters were used in the classification of all scenes of a subject, this shows that stable classifier parameters can be found for each person. However, there are partially large differences between the parameters for different persons, hence we guess that the parameters of one person might not generalize well to other people.

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	85	72	83	92	84	61	64	67	66	75	80	75.4	10.1
AAM – success	80	65	82	85	75	60	48	72	17	58	63	64.1	19.3
– failure	89	83	83	100	94	62	77	62	91	92	91	84.0	12.5
– all	81	66	80	92	88	57	66	78	57	58	73	72.4	12.5
GEF – success	93	53	86	90	88	73	44	84	0	50	58	65.4	28.1
– failure	71	83	72	94	88	38	84	69	87	67	83	76.0	15.4
– all	91	66	80	97	88	61	61	72	60	71	71	74.4	12.9
gray – success	93	76	86	95	81	73	32	84	17	58	50	67.7	25.5
– failure	91	50	72	100	94	46	84	58	83	83	86	76.8	18.0

Table 5.7.: Classification accuracies for different features where the classification was performend in a leave-one-out cross validation manner where the mean classifier parameters from the cross validations experiments were used in each case (please see Sec. 5.4.1 and Tab. 5.3). For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 5.4.1.

features	para.	grid values
AAMs	σ	0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1
	C	1, 2, 5, 10, 20, 50, 100
gray	σ	0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1
	C	50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000

Table 5.8.: Values of the SVM classifier parameters σ and C that were used in the grid searches for parameter optimization, each for AAM and raw image features, whereby each video was represented by the mean feature vector of its frames only. The parameter ranges were empirically determined by preliminary tests. Please refer to Sec. 5.4.2.

5.4.2. Classification with Mean Feature Vectors

In the previous Sec. 5.4.1, we investigated a simple majority voting over the classification results for the feature vectors of all frames of a test video. This section considers an even simpler approach: each video is represented by one feature vector only, namely the mean vector of its frames. This simple classification method yielded surprisingly good results, comparable to those of the majority voting scheme, in part even slightly better. The classification performances are summarized in the upper block of Tab. 5.9 for the mean vectors of both the AAM features (row *m-AAM*) and the raw image features (row *m-gray*); the classifier parameter values that were used in the grid searches are lised in Tab. 5.8. The best performance was achieved by the mean AAM features with an average classification accuracy of 76.0%. Likewise to the majority voting over frames, the variance is high in all cases.

We tested the stability of the involved classifier parameters also in this case, in the same way we did it above. The results are shown in the lower block of Tab. 5.9. The average classification accuracies improved slightly in all cases. These small improvements are probably due to the nature of the averaging by which the mean parameters were calculated: as the mean parameters include information from all scenes, they indirectly also used the test data, because each scene was part of all but one training sets for the parameter optimizations. In a sense, the

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
Grid search for parameter optimization prior to each training:													
– all	76	83	80	95	84	57	62	74	66	71	88	76.0	11.5
m-AAM – success	67	82	89	90	81	60	52	69	25	75	83	70.3	19.2
– failure	83	83	67	100	88	54	70	81	87	67	91	79.2	13.3
– all	82	76	72	89	81	75	57	71	54	71	81	73.5	10.5
m-gray – success	73	76	82	85	81	87	40	66	0	67	71	66.2	25.5
– failure	89	75	56	94	81	62	71	77	83	75	89	77.5	11.6
Mean parameters used in all classifications:													
– all	91	66	85	92	81	68	62	79	66	71	93	77.6	11.6
m-AAM – success	93	65	89	90	81	67	52	75	42	67	88	73.5	16.7
– failure	89	67	78	94	81	69	70	85	78	75	97	80.3	10.1
– all	91	72	74	89	88	64	59	79	63	75	85	76.3	11.2
m-gray – success	80	76	79	85	94	67	48	81	0	67	79	68.7	25.7
– failure	100	67	67	94	81	62	68	77	96	83	89	80.4	13.3

Table 5.9.: Classification accuracies for different features where the classification was performend in a leave-one-out cross validation manner either with parameter optimization on the training data prior to each training (upper block) or using the mean parameters of these optimizations (lower block). Each video is represented by one feature vector only, namely the mean feature vector of all its frames. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 5.4.2.

features	01	02	03	04	05	06	07	08	09	10	11
AAM – $\sigma \cdot 10^3$	2.80	0.29	0.15	7.57	0.07	2.70	0.10	0.22	1.74	0.92	0.05
– C	79	347	328	66	161	3	212	217	74	135	476
gray – $\sigma \cdot 10^2$	0.01	3.13	2.75	0.01	0.38	0.45	5.20	0.66	0.01	0.01	0.07
– $C \cdot 10^{-2}$	161	70	42	105	184	133	38	122	32	179	180

Table 5.10.: Mean classifier parameters σ and C that were used for the training of all scenes of a person, for AAM and raw image features, where each video was represented by the mean feature vector of its frames only. Please refer to Sec. 5.4.2.

mean parameters appear to accumulate some useful information about the feature vectors of a person which leads to a slight improvement of the classification accuracies.[5] Thus, the purpose of the classification with the mean parameters is to demonstrate that a single, stable set of parameters can be chosen for the classification of all scenes of a person. In terms of absolute classification performance, these results are not decisive, but the classification accuracies from the individual grid searches only on the training data reported above are. Table 5.10 lists the mean classifier parameters. Due to the large differences between the parameters for different people, we expect that the parameters of one person will not generalize well to other people, similar to the majority voting over frames.

[5] Although the improvements caused by the usage of the mean classifier parameters are not statistically significant, we suppose that they are systematic, because they were consistently found in almost all investigated cases.

features	01	02	03	04	05	06	07	08	09	10	11
– inter	18.4	13.8	25.4	21.0	30.3	38.5	20.5	23.2	17.6	18.2	20.0
AAM – success	12.9	14.4	24.9	20.2	32.6	39.0	19.2	25.3	19.2	18.8	18.3
– failure	16.1	11.2	23.6	12.0	22.8	38.5	20.3	15.5	13.8	17.1	11.6
– inter	1.70	1.22	1.80	1.79	1.63	1.61	1.42	1.62	0.97	1.67	1.67
gray – success	1.73	1.22	1.69	1.62	1.64	1.43	1.48	1.52	1.04	1.78	1.67
– failure	1.31	1.12	1.83	1.03	1.45	1.71	1.31	1.29	0.83	1.60	1.26

Table 5.11.: Mean intra- and interclass distances of the mean feature vectors, each for different persons and for AAM and raw image features. The *inter* rows show the mean interclass distances, i.e. the mean pairwise euclidean distances between all *success* and all *failure* frames. The *success* rows list the mean pairwise euclidean distances between all *success* frames, likewise do the rows *failure* for all *failure* frames; these are the mean intraclass distances of the respective class. Please refer to Sec. 5.4.2.

Similarly to the majority voting, the classification accuracy achieved for a subject is related to the ratio of mean interclass to intraclass distance of the feature vectors of this subject (please see Tab. 5.11). This correlation is significant for both the mean AAM features (Spearman correlation, $\rho \approx 0.79$, $p < 0.01$) and the mean raw image features (Spearman correlation, $\rho \approx 0.63$, $p < 0.04$). Again, this supports the hypothesis that a low interclass to intraclass variance ratio is a main reason for the difficulty of the classification problem.

The question arises why the mean feature vectors performed surprisingly well, compared to the performance of the majority voting over all feature vectors of a scene. A closer inspection of the majority voting results unveiled that very often there are one or two subsequences of the video where almost all frames were correctly classified, and also one or two subsequences where almost all frames were misclassified. In cases where the latter outnumbered the former ones in terms of total length, the scene was necessarily misclassified by the majority voting scheme, independent of the confidence of the single classification decisions. Thus, it seems that only (possibly short) subsequences of the videos are actually discriminative in terms of valence, although the videos are segmented to contain only the relevant part of the interaction, i.e. the reaction of the person to the robot's answer (please refer to Sec. 3.2.1). This assumption is supported by a visual inspection of the videos. Hence, in spite of the presegmentation, the videos appear to contain a significant number of frames that are irrelevant for a discrimination of *success* and *failure*, those frames are likely to disturb the majority voting. In contrast, the mean feature vectors could capture important characteristics of the associated class, even in case the irrelevant feature vectors (slightly) outnumber the discriminative ones, as the latter have still a significant influence on the value of the mean feature vector. Hence, majority voting over the complete video sequence is not well suited for a large number of scenes. Instead, an automatic detection of important subsequences appears to be a promising idea for further investigations.

5.4.3. Comparison to the Human Recognition Performance

In this section, the classification results of the best-performing features, the mean AAM feature vectors, are compared to the human recognition results, more concretely to the results for the *only-face/full-time* context condition, because it matches best the information the automatic recognition method investigated in this chapter could use (please refer to Sec. 3.5.2). For the

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	76	83	80	95	84	57	62	74	66	71	88	76.0	11.5
m-AAM – success	67	82	89	90	81	60	52	69	25	75	83	70.3	19.2
– failure	83	83	67	100	88	54	70	81	87	67	91	79.2	13.3
– all	82	75	85	92	68	73	94	67	78	95	92	82.0	19.1
human – success	91	66	84	89	61	70	91	52	66	95	93	78.1	21.2
– failure	73	84	86	95	75	75	98	82	91	95	91	86.0	16.1

Table 5.12.: Classification accuracies for the mean AAM features and the human recognition performance. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 5.4.3.

sake of convenience, these classification accuracies are shown again in Tab. 5.12.

The average classification accuracy of the mean AAM features (76.0%) is notably lower than the average human performance (82.0%), although the differences are not statistically significant.[6] However, we assume that the high variances in the performances for different persons, paired with the comparatively low number of persons, are the reason why the significance of the differences cannot be confirmed, while in fact the classification accuracies of the automatic approach are systematically lower than the human ones, not just by chance. Then again, the human recognition performance was evaluated on a subset of 88 videos only, while the automatic classification used all available videos. When evaluated on this subset of videos only, the performance of the mean AAM features is comparable to the human one: 83.0% for all videos (SD 10.1), 75.0% for *success* videos (SD 19.4), and 90.0% for *failure* videos (SD 12.6). These 88 videos were randomly chosen (please see Sec. 3.5.1). It might be the case that—by chance—these 88 videos are in some general sense "easier" to classify than the average of the database, but just as well the performance increment for the mean AAM features on this subset might be by chance; the data at hand does not allow a conclusive answer to this question (intuitively, we suspect the latter).

There are some commonalities between human and automatic recognition performances:

- on average, *failure* scenes were easier to classify than *success* scenes
- the variance for *success* scenes is higher than for *failure* scenes
- the variance of the classification accuracy (depending on the subject) is high in general

Nevertheless, there is no significant correlation at all regarding the classification accuracies for the individual persons (Spearman correlation, $\rho \approx 0.04$, $p > 0.9$). However, this question can also be considered in a more detailed way, namely not on person level, but on video level. In the latter case, the single classification results for all 88 videos are compared, while in the former one, the average classification accuracies of the 11 subjects are evaluated. In order to do this, the classification results for the 11 observing subjects[7] were binarized for each video by setting the classification result to 1 if more than half of the subjects classified it correctly, and to 0 otherwise. This binarization was done to become compatible with the results of the automatic recognition, which yielded only one binary value (correct or false classification) for each video. It turned out that there is a weak, but close to significant correlation between

[6] $p > 0.2$ for both a two-tailed t-test and a Wilcoxon rank sum test

[7] There were 44 observing subjects, who were distributed over the four context conditions, thus resulting in 11 observing subjects for each context condition, not to be confused with the 11 subjects shown in the videos.

these classification results on the 88 videos (Spearman correlation, $\rho \approx 0.2$, $p < 0.06$). Thus, measured on video level, the human observers and the automatic classification tended to make some similar classification errors to some (weak) extent.

5.5. Conclusion

We investigated the person-specific automatic recognition of FCSs in terms of valence using SVMs as classifier and AAMs, GEFs, and raw images as features. Although shown to yield good results on other facial analysis problems, the GEF features performend worse than the AAM and also raw image features in our evaluations. The good performance of the raw images, compared to the AAMs, suggests that also the video parts with large out-of-plane head rotations, which are a main cause for AAM fitting failures, convey useful information and should be considered for the interpretation. In general, the achieved classification accuracies are rather low for a two-class problem, espescially for the *success* class. A main problem is the apparently low interclass to intraclass variance ratio on frame level.

The best performance was achieved by the mean AAM feature vectors, yielding an average classification accuracy of 76.0%, which is still lower than the average human performance of 82.0%. When evaluated only on the subset of videos that was judged by the human subjects, the classification accuracy increased to 83.0%. However, we regard the classification performance for the whole dataset as the more important performance measure. Likewise to the human classification, the variances of the recognition performances for different persons were very high in general and for *success* scenes in particular. On average, *failure* scenes were somewhat easier to classify than *success* scenes.

An investigation of the surprisingly good performance of the mean feature vectors, compared to the majority voting over frames, indicated that the usage of descriminative subsequences of the videos for the classification appears to be a promising direction for further investigations. This assumption is confirmed by a visual inspection of the videos, which furthermore suggests that the temporal dynamics of the displayed FCSs are important for their recognition. Both issues were neglected by the simple static classification approach presented in this chapter. Thus, we investigate a more sophisticated and dynamic recognition approach that addreses them in the next chapter. The classification accuracies of the static approach serve as baseline for the dynamic approach to compare to.

6. A Dynamic Recognition Approach

Ideas aren't magical; the only tricky part is holding on
to one long enough to get it written down.
— Lynn Abbey

This chapter presents our dynamic recognition approach for the classification of FCSs in terms of valence, where "dynamic" refers to the consideration of the temporal dynamics in the facial displays. It is generally assumed that these dynamics are important to accurately interpret spontaneous, authentic FCSs, this was already noted long ago [174]. Ekman [136] emphasized the important role of the temporal dynamics in case of spontaneous facial displays, in spite of using still images in his own research on posed facial expressions (please refer to Sec. 2.2.3). Also Bruce *et al.* [46] discussed the dynamic nature of facial expressions and suggested that humans might combine a static and a dynamic representation in their recognition processes. Furthermore, a visual inspection of the object-teaching videos also suggests an important role of the temporal dynamics of the shown facial displays.

The dynamic recognition approach is based on the selection of (comparatively short) discriminative subsequences of the input videos which serve as prototypes of the respective class. This is motivated by the observation reported in Sec.5.4.2 suggesting that very often only a short subsequence of the object-teaching videos is actually discriminative in terms of *success* and *failure*; this is also supported by a visual inspection of these videos. The temporal dynamics are considered by means of dynamic time warping (DTW) [433] which provides an elastic distance measure between subsequences. This choice of a distance measure is motivated by the results of Ding *et al.* [115]. They performed comprehensive experiments with several different distance functions on a large number of time series datasets from various application domains, where the DTW distance yielded very good results in almost all cases.

The classification of new data is done by a nearest-neighbor-based (NN) classification technique. Despite being simple in their structure, NN classifiers have shown good performance in several time series classification problems. Xi *et al.* [526] conducted a series of experiments where they demonstrated that a simple NN classifier—combined with DTW as distance measure—outperforms several more sophisticated classification approaches on various datasets.

The remainder of this chapter is organized as follows. Section 6.1 explains our dynamic recognition approach in detail. Other approaches that are closely related are discussed in Sec. 6.2. An evaluation of the dynamic recognition approach on the videos of the object-teaching scenario (please see Sec. 3.2) is presented in Sec. 6.3. The last Sec. 6.4 concludes this chapter with a critical review of the achieved results.

6.1. Classification based on Reference Subsequences

This section explains our dynamic approach for FCS recognition. It essentially involves the search for comparatively short video subsequences with high discriminative power. These subsequences are used as prototypical representatives for the two classes in a classification technique considering temporal dynamics. A subsequence is a set of several consecutive frames of a video, where each frame is represented by the corresponding AAM parameter vector. The "discriminative power" of a subsequence refers to its suitability to distinguish *success* from *failure* videos (please see below). The presented approach consists of the following major steps, which are explained in detail in the subsequent sections:

1. For all possible subsequences (within a certain range of length) of all videos of the given training data, a "discriminativity"-value is computed. This value is high for subsequences that are similar to other subsequences of the same class, but are rather different to any subsequence of the opposite class. Thus, a high discriminativity-value indicates a subsequence with high discriminative power. To account for the temporal nature of the subsequences, dynamic time warping (DTW) [433] is used as distance measure between subsequences. [→Sec. 6.1.1]
2. From all considered subsequences, a certain number of subsequences with high discriminativity-values is chosen as reference subsequences for each class. [→Sec. 6.1.2]
3. These reference subsequences are used as prototypes in a nearest-neighbor-based classification. [→Sec. 6.1.3] To take into account the possibly different expressiveness of a person regarding positive and negative FCSs, this classification scheme is extended by introducing a bias that favors one class over the other. [→Sec. 6.1.4]
4. This classification approach involves several parameters which are optimized on the training data by means of model selection techniques. Therefore, the steps 1. to 3. are iterated over different parameter sets to perform a leave-one-out cross-validation on the training data for parameter optimization. [→Sec. 6.1.5]

Section 6.1.6 finally outlines the basic implementation design that we used for our experiments.

6.1.1. Discriminative Subsequence Detection

The goal of the discriminative subsequence detection is to find (comparatively short) video subsequences within the input videos that are characteristic for either *success* or *failure* scenes and can thus be used as prototypical reference subsequences to classify a new scene. Each video is represented as a sequence $A = a_1 a_2 \dots a_N$ of AAM frame parameter vectors a_i of the face, normalized to zero mean and unit variance. Such a normalization is required to suitably compute the similarity of two sequences [281]. In order to find suitable subsequences, an exhaustive search over all possible subsequences of length $l \in [l_{\min}, l_{\max}]$ (in frames) of all training video sequences is performed.

For each subsequence $x_{m,i}$ of each video, a discriminativity-value $s_{m,i}$ is computed:

$$s_{m,i} = \frac{\sum k\_\min_{n,j} \{\, d_m^n(i,j) \mid c_m \neq c_n, \qquad j \in P_{m,i}^n \,\}}{\sum k\_\min_{n,j} \{\, d_m^n(i,j) \mid c_m = c_n, n \neq m, j \in P_{m,i}^n \,\}}, \tag{6.1}$$

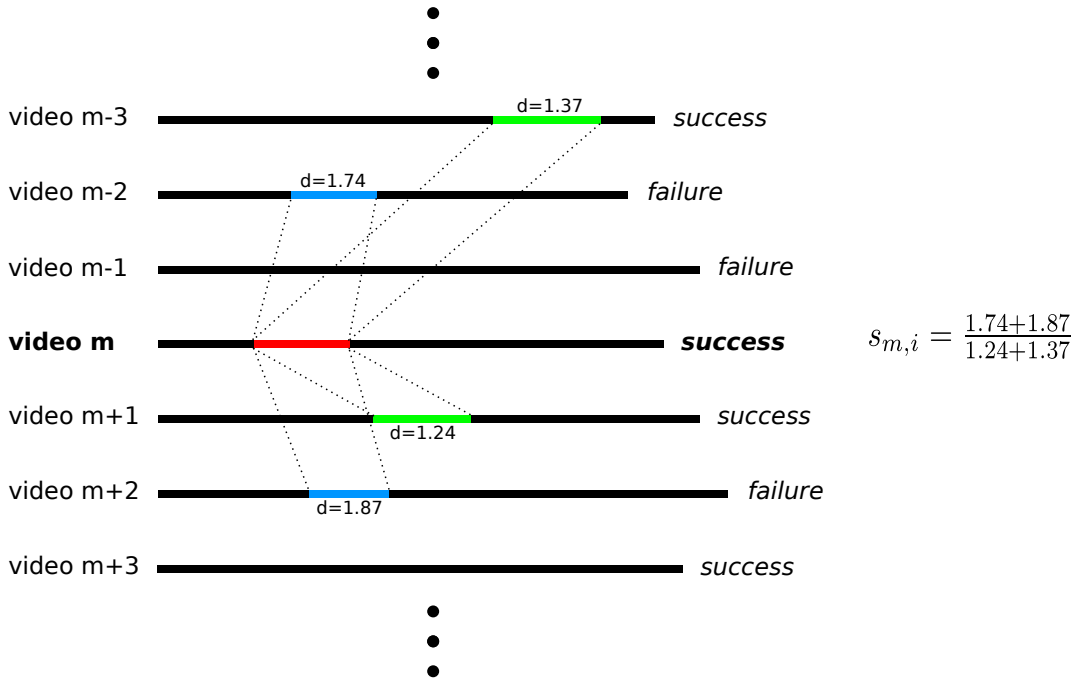


Figure 6.1.: Example depiction of the discriminativity-value computation defined by Eq. 6.1. It illustrates the computation of $s_{m,i}$ for the i-th subsequence (shown in red) of the m-th video for $k = 2$. In search of those subsequences that have minimal distance to this target subsequence, the subsequences of all other videos are considered (applying the length constraint in Eq. 6.2). Concerning the videos of the same class as the m-th video, *success*, the two subsequences of the $(m + 1)$-th and $(m-3)$-th video (shown in green) are found to have minimal distances (1.24 resp. 1.37) to the target subsequence. Similarly, the two subsequences of the $(m - 2)$-th and $(m + 2)$-th video (shown in blue) have the minimal distances (1.74 resp. 1.87) of the subsequences from all videos of the opposite class, *failure*. Hence $s_{m,i} = \frac{1.74+1.87}{1.24+1.37}$ according to Eq. 6.1. Please refer to Sec. 6.1.1.

where m and i are the indices of the i-th subsequence in the m-th video, $k\_\min\{X\}$ denotes the k smallest values of set X, $d_m^n(i, j)$ is the normalized distance of subsequence $x_{m,i}$ (the i-th subsequence in the m-th video) to subsequence $x_{n,j}$, c_m denotes the class (*success* or *failure*) of the m-th video, and $P_{m,i}^n$ is the index set of all subsequences in the n-th video, the lengths of which are constrained by the length of $x_{m,i}$:

$$P_{m,i}^n = \{\ j\ \mid \lfloor l_{m,i}/f \rceil \leq l_{n,j} \leq \lfloor l_{m,i} \cdot f \rceil\ \mid j \in M_n\ \}, \tag{6.2}$$

where $l_{m,i}$ is the length (in frames) of subsequence $x_{m,i}$, M_n is the index set of all subsequences in the n-th video, and $f \geq 1$ is a factor describing the maximum allowed difference in length of two subsequences. Thus, $x_{m,i}$ is not compared to all subsequences of all other videos, but to a subset of these subsequences, namely those that do not differ in length too much. This avoids comparison of subsequences of very different lengths and thus prunes the search space for the calculation of $s_{m,i}$, which is the sole purpose of this constraint.

In all experiments described later in this chapter, $f = 1.3$ was pragmatically chosen based on

some preliminary tests, as this value is expected to be a reasonable compromise between evaluating all relevant subsequences and pruning the search space to avoid needless computations. Values significantly higher are not expected to influence the resulting discriminativity-value $s_{m,i}$, as according to Eq. 6.1 only the k smallest distances are considered, and two subsequences with very different lengths are unlikely to have a small distance to each other, thus it seems safe to drop those comparisons. Nevertheless, a high f-value would substantially increase the computational effort because many irrelevant distances needed to be calculated. On the other hand, f should not be chosen too small to avoid the undesired pruning of some relevant subsequences.[1]

The distance $d_m^n(i,j)$ of two subsequences is computed via dynamic time warping (DTW) [433] over the AAM parameter vector sequences. The resulting distance value is normalized by the length $l_{m,i}$ to allow for fair comparison of subsequences of different lengths in Eq. 6.1.

Equation 6.1 yields high discriminativity-values for subsequences with low minimal distances to subsequences of videos representing the same class (denominator) and high minimal distances to subsequences of videos representing the opposite class (numerator). In other words, the discriminativity-value is high for subsequences that are very similar to other subsequences of the same class and at the same time rather different from even the most similar subsequences of the opposite class. This is similar to the Fisher criterion [164], which minimizes the within scatter while maximizing the between scatter of data from two classes to find an optimal discriminant function. Thus, the higher the discriminativity-value of a subsequence (compared to the discriminativity-values of other subsequences of the given video set), the better it is suited as a representative of the respective class for discrimination purposes. Figure 6.1 shows an example illustration of the discriminativity-value computation.

6.1.2. Reference Subsequence Selection

For each of the two classes, t non-overlapping subsequences with high discriminativity-values are selected as reference subsequences. It might be beneficial for the classification to not select the t subsequences with the t highest discriminativity-values overall, but to preferably select v subsequences per video, for the following reason: If a small number of videos of one class c is very similar to each other and also rather different to any video of the other class, the major part of the t subsequences with highest discriminativity-values overall might stem from these few videos. A larger number of videos of class c might be typical for this class as well, but not that similar to the aforementioned small group of videos. This larger group would be underrepresented by the reference subsequence selection. Thus, the resulting classifier would be able to classify videos similar to the small group very confidently, but would probably perform poor for videos similar to the larger group. To avoid this problem, a more uniform distribution of reference subsequences over the training videos is required. This motivates the following selection method.

S_c is the index set of the v non-overlapping subsequences with the highest discriminativity-values for each training video of class c:

$$S_c = \bigcup_{m|c_m=c} \{\ (m,i)\ \mid i \in R_m^v\ \}, \tag{6.3}$$

[1] Please refer to Sec. 6.2.1 for a discussion why it is difficult to utilize more sophisticated pruning schemes.

where $R_m^v \subseteq M_m$ is the index set of the v non-overlapping subsequences with the highest discriminativity-values in the m-th video. Further, Q_c contains $(t - v)$ of the remaining non-overlapping subsequences with the highest discriminativity-values per video that are not part of S_C:

$$Q_c = \bigcup_{m|c_m=c} \{ (m,i) \mid i \in R_m^t \, , \, (m,i) \notin S_c \}. \tag{6.4}$$

The index set R_c of the t final reference subsequences for class c is given by a combination of the elements of S_c and Q_c:

$$R_c = \begin{cases} t\underset{(m,i)}{\text{arg max}} \{ s_{m,i} \mid (m,i) \in S_c \} & \mid \text{if } \|S_c\| \geq t \\ S_c \quad \cup \quad T_c & \mid \text{if } \|S_c\| < t \end{cases} \tag{6.5}$$

$$\text{with} \quad T_c = (t - \|S_c\|)\underset{(m,i)}{\text{arg max}} \{ s_{m,i} \mid (m,i) \in Q_c \}, \tag{6.6}$$

where $t\text{arg max}\{X\}$ denotes the arguments associated with the t largest values of the set X.

In summary, for each video of class c, the v (non-overlapping) subsequences with the highest discriminativity-values are determined and (the indices of them) collected in S_c (Eq. 6.3). The t best of these subsequences are chosen as reference subsequences of class c. In case S_c contains less than t elements, the missing ones are taken from the best remaining (non-overlapping) subsequences that are not part of S_c, i.e. from T_c (Eq. 6.4 and 6.6). Hence the index set R_c of the t reference subsequences for class c is complete (Eq. 6.5).

6.1.3. Nearest-Neighbor-based Classification

The classification of a test video sequence (index m) starts with the computation of the minimum distance $d^*_{m,(n,j)}$ of every reference subsequence $x_{n,j} \in R_{\text{success}} \cup R_{\text{failure}}$ to all subsequences (index i) of the test video, considering a similar pruning condition for the involved subsequence lengths as in 6.2:

$$d^*_{m,(n,j)} = \min \{ d^n_m(i,j) \mid i \in M_m \}. \tag{6.7}$$

For each class c, the u smallest distances are combined to get a classification score $d_{m,c}$:

$$d_{m,c} = \sum_{\gamma \in \Gamma} \frac{1}{\gamma^w} \quad , \; \Gamma = u\text{min} \{ d^*_{m,(n,j)} \mid (n,j) \in R_c \}, \tag{6.8}$$

where the parameter w weights the influence of large distances compared to small ones. The test video sequence is classified into the class c^* with the highest classification score d_{m,c^*}. This is a k-nearest-neighbor-based classification approach (k-NN), as the best distances to a certain number of reference subsequences are combined to form the final classification decision.

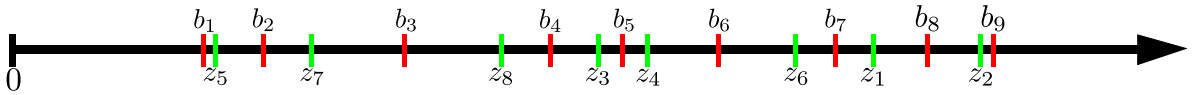


Figure 6.2.: Example depiction of the classification bias computation. The quotients of the classification scores $z_g = \frac{d_{g,\text{success}}}{d_{g,\text{failure}}}$ (shown in green) for training videos g are significant for the computation of the classification bias candidate values b_i (shown in red). Each bias b from the interval formed by two neighboring z_g values leads to the same classification result on the training data, thus only one value from this interval needs to be considered. Hence the mean value of the interval borders is chosen due to its maximum margin (e.g. b_4 for (z_8, z_3). Completed by two marginal values (b_1 and b_9), these mean values constitute the set of bias candidate values that are evaluated for their classification accuracy on the training data. Please refer to Sec. 6.1.4.

6.1.4. Biased Classification

The degree of expressiveness of positive compared to negative valence might vary considerably, depending on the individual characteristics of a person. While some people display both with approximately the same expressiveness, others might show a clear bias, meaning that the absence of *failure* signs could reasonably be interpreted as *success*, or vice versa. Taking this into account, we introduce a bias b on the classification scores:

$$d'_{m,\text{success}} = d_{m,\text{success}} \quad , \quad d'_{m,\text{failure}} = b \cdot d_{m,\text{failure}}, \tag{6.9}$$

where d'_c is the new classification score for class c. The value of b is chosen such that the training error is minimized. The candidate values for b are computed as follows: For each classification of a training video (index g), the quotient z_g of the classification scores is calculated:

$$z_g = \frac{d_{g,\text{success}}}{d_{g,\text{failure}}}. \tag{6.10}$$

As $d'_{g,\text{success}} = d'_{g,\text{failure}}$ holds for $b = z_g$, these z_g values are the points where changes in the classification results of the training data occur when one alters b. Thus, when one sorts all z_g values in ascending order, for any two neighboring values z_{g_1} and z_{g_2}, all selections of b from the interval (z_{g_1}, z_{g_2}) will yield the same classification result, hence only one value $b \in (z_{g_1}, z_{g_2})$ needs to be considered as candidate for the optimization. We choose the mean values of the interval borders in each case (because they have maximum margin to the "change points" for the classification and thus are reasonable choices with respect to generalization). Together with one value slightly below the minimum z_g value and another value slightly above the maximum z_g value, they constitute the candidate values for the optimization. Finally, we select the value $b = b^*$ that yields the best classification result on the training data. If there are several best values, the median of them is chosen. Figure 6.2 gives an example illustration of this bias computation.

parameter / description	grid values
$[l_{\min}, l_{\max}]$: considered subsequence lengths [→ Sec. 6.1.1]	[5,5], [10,10], [15,15], [5,20]
k: number of distances for subsequence scores [→ Eq. 6.1]	1, 2, 5, 10, 15
t: number of reference subsequences in total [→ Sec. 6.1.2]	1, 2, 5, 10, 15, 20, 25
v: number of reference subsequences per video [→ Eq. 6.3]	0, 1, 2
u: number of distances for classification scores [→ Eq. 6.8]	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
w: distance weight [→ Eq. 6.8]	1, 2, 3
b: classification bias [→ Sec. 6.1.4]	1.0, b^*

Table 6.1.: Overview of all parameters of the dynamic recognition approach for FCSs classification described in Sec. 6.1. The left column shows the parameters together with the section describing them resp. the equation where they are involved. The right column lists the candidate values of these parameters that were used in the grid search for parameter optimization in the experiments reported in Sec. 6.3. Parameters that influence the training are listed in the upper block, those mainly affecting the classification of test data in the lower one. However, as the training involves trial classifications of training data for parameter optimization, these classification parameter also influence the training indirectly. Please refer to Sec. 6.1.5.

6.1.5. Parameter Optimization

This classification approach involves several parameters that need to be set. They are optimized on the training data by means of a grid search over different candidate parameter sets, where a leave-one-out cross-validation is performed for each set to test its suitability: For all possible combinations of parameters, each training video is treated as test data once, whereas all remaining videos are used to train the classifier.[2] Finally, the parameter set yielding the best classification rate is selected and used to train the classifier on all training videos. In case of several parameter sets showing the same optimal performance, the set with the highest margin ψ is selected:

$$\psi = \frac{\sum_{m|r_m=c_m} |d'_{m,\text{success}} - d'_{m,\text{failure}}|}{\sum_{m|r_m\neq c_m} |d'_{m,\text{success}} - d'_{m,\text{failure}}|}, \tag{6.11}$$

where r_m is the classification result for the m-th video. This auxiliary value ψ is high for correctly classified videos with a high difference in classification scores ("confidently correct") and for misclassified videos with a low difference in classification scores ("near miss"). Thus, this parameter selection tries to improve generalization by choosing those parameters that led to the safest classification result on the training data.

A complete list of all parameters together with their values used in the grid search in the experiments reported in Sec. 6.3 is given in Tab. 6.1.

6.1.6. Implementation

We did a prototype implementation of the dynamic recognition approach in Matlab. As this approach involves an exhaustive search for suitable reference subsequences during the classifier

[2] During this grid search, some constraints regarding valid parameter combinations are applied, for instance $v \leq t$ and $u \leq t$ need to hold.

training, it was important to pay attention to optimization possibilities in order to keep the training tractable regarding its runtime. The implementation is outlined in App. A.6 and the algorithms Alg. A.1 to Alg. A.12 listed there. Here, we briefly discuss the basic ideas behind.

The model training—the selection of reference subsequences—is outlined in Alg. A.1. It starts by precomputing the Euclidean distances of the AAM parameter vectors of all frames of a training video to those of all frames of all other training videos (Alg. A.4). These distances are frequently required in the remainder of the training process. This precomputation requires $O(N^2 \cdot L^2 \cdot D)$ operations and $O(N^2 \cdot L^2)$ space,[3] where N is the number of training videos, L is the average length (in frames) of a training video, and D the dimension of the AAM parameter vectors.[4] Hence the retrieval of frame distances is a table lookup of $O(1)$ subsequently.

To support the discriminativity-score calculations, a large matrix of certain minimal subsequence distances is precomputed (Alg. A.6). This matrix allows $O(1)$ access to the K minimal distances of the subsequence of the m_1-th video that starts at frame s_1 and is l_1 frames long to any subsequences of the m_2-th video; for any valid values of m_1, s_1, l_1, and m_2. The indexing regarding m_1, s_1, and l_1 is necessary to determine the minimal distances of a particular subsequence (m_1, s_1, l_1) which is required in the discriminativity-score calculation (please refer to Eq. 6.1). The additional indexing regarding m_2 is beneficial because during parameter optimization, a leave-one-out classification of each training video is performed. To be able to exclude each video from the training data once, one needs to ensure that none of the minimal distances of a candidate subsequence used in Eq. 6.1 is in fact a distance to this excluded video, because it is considered unknown test data in this situation. This is accomplished by this indexing regarding m_2 which explicitly states whose video's subsequences are considered.

The precomputation of this subsequence distance matrix takes $O(N^2 \cdot L^2 \cdot (M_l^2 \cdot M_k + L^2))$ operations, where M_l is the maximal length of the reference subsequences. evaluated in the parameter grid search and M_k is the maximal number of minimal distances to consider in the discriminativity-score calculations (please see Tab. 6.1). This includes the computation of several DTW matrices (Alg. A.5). Commonly, $L^2 > M_l^2 \cdot M_k$ is expected, in which case the runtime becomes $O(N^2 \cdot L^4)$. However, it is possible to parallelize the first four loops of Alg. A.6 which would result in $O(N^2 \cdot L^2)$ parallel computations of $O(L^2)$ runtime each in the extreme case. For the evaluations presented in Sec. 6.3, we parallelized the first loop in some cases, thus performing $O(N)$ parallel computations of $O(N \cdot L^4)$ runtime each. The computed subsequence distance matrix requires $O(N^2 \cdot L \cdot M_l \cdot M_k)$ space.[5]

Aided by the precomputed frame and subsequence distance matrices, the grid search for parameter optimization is performed (Alg. A.3). In an outer parameter loop, all combinations of $(l_{\min}, l_{\max})$ and k parameters are evaluated. These parameters fundamentally affect the selected reference subsequences, therefore they need to be recomputed in each iteration. Inside this loop, a trial classification of each training video is performed. This involves the selection of a certain number of reference subsequences per training video (excluding the respective test video of the trial classification) (Alg. A.8) and the subsequent fusion of them to form the final set of reference subsequences (Alg. A.9). The selection algorithm (Alg. A.8) makes use of the

[3] This results in about 33 MB per person on average for the evaluations presented in Sec. 6.3 when stored as binary MAT-file.

[4] In principle, it is possible to parallelize these calculations as they are mostly independent, resulting in $O(N^2 \cdot L^2)$ parallel calculations of $O(D)$ runtime in the extreme case. However, this was not used and did not appear to be necessary for the evaluations presented later in this chapter.

[5] This results in about 137 MB per person on average for the evaluations presented in Sec. 6.3 when stored as binary MAT-file.

precomputed subsequence distances to calculate the discriminativity-scores. The separation in video-specific selection and subsequent fusion of reference subsequences allows for an efficient intermixture of the grid search over v, t, u, and w parameters with the reference subsequence computation process (please refer to Alg. A.3). After this is done, the smallest distances of any subsequences of the trial test video to these reference subsequences are determined (Alg. A.7), the resulting classification scores are computed (Alg. A.10), and the classification rate and margin is calculated (Alg. A.12) for both an unbiased and a biased classification (Sec. 6.1.4, Alg. A.11). Finally, the parameter set that yielded the best classification rate (and best margin in case there are several parameter sets with this classification rate) is chosen as parameter set. The runtime of the whole grid search heavily depends on the grid values that are evaluated:

$$O(\, n_l \cdot n_k \;\cdot\; (\, N \cdot (\, L \cdot M_l \cdot (N \cdot M_k + M_t) \;+\; n_v \cdot M_t \cdot (N \cdot M_v + L \cdot M_l^2) \;+ \\ n_t \cdot n_u \cdot n_w \cdot (M_u + M_t \cdot \lg M_t)\,) \;+\; n_v \cdot n_t \cdot n_u \cdot n_w \cdot N \cdot \lg N\,)\,),$$

where n_X is the number of grid values for parameter X and M_X is the maximum value of parameter X. When common relations between the involved variables are considered, this simplifies to $O(n_l \cdot n_k \cdot n_v \cdot M_t \cdot M_l^2 \cdot N^2 \cdot L)$ which still reflects the strong influence of the grid values (which is sane due to the nature of the grid search). The outer parameter loop discussed above can be parallelized, in which case $O(n_l \cdot n_k)$ parallel computation processes of $O(n_v \cdot M_t \cdot M_l^2 \cdot N^2 \cdot L)$ runtime each would be utilized, where each process evaluates the performance of a particular $(l_{\min}, l_{\max}, k)$ parameter combination and the best of these results is chosen afterwards. A further parallelization of the trial classification loop is also possible (resulting in $O(n_l \cdot n_k \cdot N)$ parallel processes of $O(n_v \cdot M_t \cdot M_l^2 \cdot N \cdot L)$ runtime each) at the cost of a more complex final evaluation. The memory consumption of the grid search is about $O(n_v \cdot n_t \cdot n_u \cdot n_w + n_l \cdot n_k + N \cdot \lg N + M_t \cdot (M_v + M_l^2))$.

The training is completed by the final selection of reference subsequences using the best parameters found in the grid search (Alg. A.1). When one regards the dimension of the AAM parameter vectors and the parameter grid values as constant and focuses on the influence of the training data only, the runtime is $O(N^2 \cdot L^4)$, dominated by the time used for the subsequence distance precomputation. In practice, both this precomputation and the parameter grid search usually take a comparable large amount of time. For completeness, the runtimes of the various auxiliary algorithms mentioned above are summarized in Tab. 6.2. Due to the strong influence of the maximum video length L on the runtime, the dynamic recognition approach is applicable for comparatively short video sequences (like the object teaching scenes investigated in this work) only. Clearly, the training is to be performed offline: for the evaluations presented later in this chapter, typical (non-parallelized) runtimes range from several hours up to a few days.

The classification of a test video is outlined in Alg. A.2. It is very similar to the trial classifications during the grid search. The minimal distances of any subsequences of the test video to the reference subsequences are computed. Based on these distances, the classification scores are calculated and the video is classified into the class with higher classification scores, considering the bias and the other parameters determined in the training. This classification requires $O(M_t \cdot M_l^2 \cdot L_t + M_u + M_t \cdot \lg M_t)$ operations and $O(M_u + M_t \cdot \lg M_t + M_l^2)$ space, thus the runtime is dominated by the length of the test video L_t and the maximum length of the reference subsequences M_l. Not only asymptotically, but also regarding the concrete

description	algorithm	runtime	space
dynamic time warping	Alg. A.5	$O(L^2)$	$O(L^2)$
compute minimal distances	Alg. A.7	$O(M_t \cdot M_l^2 \cdot L)$	$O(M_t + M_l^2)$
select reference subsequences	Alg. A.8	$O(L \cdot M_l \cdot (M_k \cdot N + M_t))$	$O(M_t)$
fuse reference subsequences	Alg. A.9	$O(N \cdot M_v \cdot M_t)$	$O(N \cdot M_v \cdot M_t)$
compute classification scores	Alg. A.10	$O(M_u + M_t \cdot \lg M_t)$	$O(M_u + M_t \cdot \lg M_t)$
compute classification bias	Alg. A.11	$O(N \cdot \lg N)$	$O(N \cdot \lg N)$
comp. classification rate/margin	Alg. A.12	$O(N)$	$O(N)$

Table 6.2.: Runtime and memory consumption for the auxiliary algorithms discussed in Sec. 6.1.6. Please refer to Sec. 6.1.6 for details on the involved variables.

number of operations (corresponding to the constant terms that are usually omitted in big O notation), the classification is much less demanding than the training. Hence it can be performed online in soft real-time for typical values of L_t.

6.2. Related Approaches

Most research on the detection of specific subsequences in sequential input data has been conducted in the data mining community. Tiwari *et al.* [488] presented a survey on methods to find frequently occurring patterns in large datasets. In typical state of the art data mining techniques for discriminative subsequence detection (e.g. [253]) and related pattern matching problems (e.g. [165]), the data are usually sequences of ordinal, univariate items (alphabets). This allows for effective pruning strategies as integral parts of the respective methods where large parts of the considered search trees can be discarded as they cannot contain a desired subsequence. Due to the continuous, multivariate feature vector data, the traits of the DTW distance measure, and other properties of our dynamic recognition approach, those pruning strategies are not applicable in our case.

For instance, Ji *et al.* [253] described an efficient method to find *minimal distinguishing subsequences* (MDSs) that occur frequently in sequences of one class, but rarely in sequences of another class. They discover MDSs by a tree search where candidate subsequences are successively extended while several pruning strategies are applied to speedup the search. Most of these strategies rely on the alphabet-property of the sequences and cannot be directly applied to sequences of multivariate data. The same holds for the approach of Floratou *et al.* [165] who presented an efficient algorithm to find all frequent "approximate" patterns in a dataset.[6] Geurts [187] suggested a different method to detect local patterns where the problem is made tractable by using a piecewise constant approximation of the time series and random sampling instead of an evaluation of the full search space. We discuss a related approach that finds an optimal solution for a similar problem in Sec. 6.2.1 in detail. Despite utilizing a decision tree in his work as preferred classifier, Geurts [187] empirically confirms the good performance of a NN classifier.

Tiwari *et al.* [488] pointed out that most pattern mining approaches focus on the performant computation of frequent patterns, leaving the quality assessment for a specific use case for

[6] Floratou *et al.* [165] point out the possibility to apply their approach to numeric data by discretizing it into symbolic sequence data like it was done in related work [394, 78]. However, these papers consider time series of univariate data. We expect that a suitable discretization is more difficult in the multivariate case where probably a substantially larger number of symbols would be needed in most cases.

subsequent processing steps. Our approach does not search for frequent patterns first, but directly tries to estimate the quality of the considered subsequences in terms of expected discrimination power.

The following sections consider four related approaches in greater detail. Finding discriminative subsequences in time series by means of *shapelets* and an extension of this technique called *logical shapelets* are discussed in Sec. 6.2.1 and Sec. 6.2.2, respectively. The classification approach of Nowozin *et al.* [380], which is also based on discriminative subsequences, is considered in Sec. 6.2.3. Buenaposada *et al.* [53] developed a classification technique for facial expressions in video streams by means of an k-NN classifier, which we discuss in Sec. 6.2.4. Finally, Sec. 6.2.5 concludes this discussion of related work.

6.2.1. Shapelets

Ye and Keogh [538, 539] presented an approach that is similar to ours in some respects. They introduced *shapelets* as new primitive for time series data mining. A shapelet is a (small) subsequence of a time series that is well suited for the discrimination of two classes. To find a shapelet, a large number of subsequences from the given training data are considered as candidates and tested for the information gain they produce when the training data set is split into two sets based on the minimal distances of subsequences from these two sets to the shapelet candidate. The information gain is defined as the difference of the total entropies before and after splitting. The shapelet is the subsequence that maximizes this information gain, together with the optimal split point (the distance threshold based on which the training time series are assigned to one of the two sets). To classify test data, shapelets are incorporated into a decision tree. At the root node, the minimal distance of any of the test time series' subsequences to the shapelet associated with this tree is calculated. In case it is less than the optimal split point, the left subtree is processed, the right subtree otherwise. This is recursively repeated until a leaf node is reached, where the test time series is classified into the class of this leaf (i.e. one training time series).

There are some similarities of Ye and Keogh's and our approach, first of all the basic idea to use prototypical reference subsequences, gained from the training data, for classification. Also the exhaustive search for these prototypes is performed in a brute force manner in both approaches (but see below for pruning strategies in Ye and Keogh's approach). Furthermore, the way Ye and Keogh compute the optimal split point is basically identical with our computation of the classification bias; both methods follow the same rationale. Despite these similarities, there are also several important differences:

- they consider real-valued time series only, whereas we use time series of multivariate data
- they use the Euclidean distance as similarity measure, whereas we utilize dynamic time warping
- they evaluate an information gain computed as entropy difference to judge the suitability of a shapelet with optimal split point, while we use the discriminativity-value (combined with the classification bias) for this purpose
- they utilize a decision tree classifier, while we use a k-nearest-neighbor-based classification approach
- they apply two effective pruning strategies which reduce the processing time considerably, whereas we use the optimizations outlined in Sec. 6.1.6, because those pruning

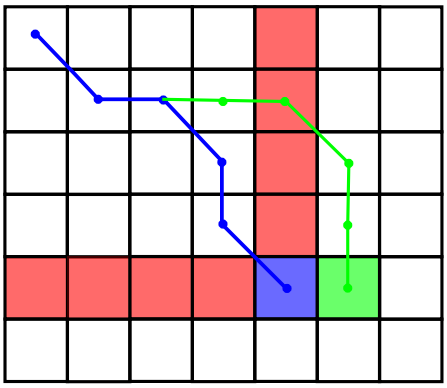


Figure 6.3.: Example illustration of a hypothetical pruning strategy for early abandon of the DTW distance computation. Suppose the DTW matrix have been constructed up to the blue element. To decide whether the computation can be abandoned at this point, it is not sufficient to compare this element to the best distance known so far, also the matrix elements shown in red need to be considered. When one moves along to the green element, the optimal alignment path might change, for instance from the blue one to the green one, hence the green element might represent a lower distance than the blue one. This possibility can be ruled out only if all red elements are larger than the best distance known so far, too. The test for this would require a runtime linear in the length of the subsequences that are compared. Please refer to Sec. 6.2.1.

strategies are not applicable in our case (please see below)

Furthermore, the kind of data Ye and Keogh used for their evaluations is different from the face data we use in this work. They investigated several datasets where they analyzed contours to classify object images (leaves, projectile points, heraldic shields), motion trajectories (Gun/NoGun dataset), spectrographic data (wheat, coffee), and artificial datasets (Lightning EMP Classification, Mallat dataset) [538, 539], but no facial data.

First Pruning Strategy: Subsequence Distance Early Abandon

The shapelet approach involves a large number of distance calculations between a shapelet candidate and all subsequences of a time series. As only the minimum distance to these subsequences is required, a distance calculation is abandoned once the current distance value exceeds the best distance known so far; the remainder of the data points in the candidate and the respective subsequence do not need to be processed because the distance can only increase further. Ye and Keogh [538] reported a halving of the processing time due to this pruning.

Unfortunately, a comparable positive effect on the runtime cannot be expected for our approach. Because shapelets use the Euclidean distance as similarity measure, the termination condition can be tested in constant time: it is just a comparison of the current distance value accumulated so far and the best distance known at this point. In case of our dynamic recognition approach, a similar test would require a runtime linear in the length of the involved subsequences due to the utilization of dynamic time warping instead of the Euclidean distance; please refer to Fig. 6.3 for an explanation. As this test needed to be performed at

several points during DTW matrix construction, it is very likely that the thereby introduced overhead annihilates the advantage of early abandon.

However, Keogh and Ratanamahatana [282, 280] presented a different pruning strategy especially for DTW. They suggested to use a specific lower bound on the DTW distance of two time series (called *LB_Keogh*) that can be computed in linear time in advance, based on reasonable constraints of the warping path. In case this lower bound is above the best distance encountered so far, the DTW matrix construction can be skipped right from the start. While originally presented for real-valued time series only [280], this concept was extended to the multivariate case by Rath and Manmatha [413]. Although this lower bound could also be utilized in the dynamic recognition approach presented here, it is not expected to yield a significant performance gain for the training, for the following reason. During the precomputation of the subsequence distances (Alg. A.6), the same DTW matrix is used for the evaluation of several subsequences, namely for all subsequences with the same starting point (and different endpoints).[7] In order to be effective, the lower bounds for the distances of *all* these subsequences have to be above the respective k best distances encountered so far (please refer to Eq. 6.1), otherwise the DTW matrix has to be constructed anyway, at least up to the maximum lengths of those subsequences where the lower bound was below the aforementioned best distances. Thus, we do not expect a high number of cases where the DTW matrix construction can actually be skipped, which limits the speedup that could possibly be gained from applying this lower bound.

For the classification of test data, the situation is slightly better in the sense that only one endpoint varies (the one of the test data subsequence as the reference subsequences are fixed, please refer to Alg. A.7). Thus the number of subsequence distances that are calculated based on the same DTW matrix is smaller compared to the training, hence the probability of skipping the DTW matrix construction is higher in this case. However, in order to avoid further complication of the software which is complex already, we did not incorporate this lower bounding strategy into our prototype implementation, thus an empirical suitability evaluation of this lower bounding strategy in our approach remains for future work.

Second Pruning Strategy: Early Entropy Pruning

Ye and Keogh [538] presented a second pruning strategy that discards a large number of distance computations, leading to a significant search space reduction "by over two orders of magnitude." [538, p. 953]. The basic idea is to abandon the evaluation of a shapelet candidate in case the distances to the training times series computed so far make it evident that this candidate cannot be better than the best candidate encountered so far. This is realized by computing an upper bound of the information gain the candidate might achieve, under the most optimistic assumptions about the not yet computed distances to the remaining times series objects. If this upper bound is below the information gain of the best shapelet candidate known so far, the evaluation of the current candidate is terminated.

Unfortunately, a similar pruning strategy cannot easily be utilized in our dynamic recognition approach. Here the criterion for judging the quality of a candidate is not the information

[7] If the starting point of one of the two subsequences to compare changes, the DTW matrix needs to be recalculated. In contrast, if only the endpoints change while both starting points stay the same, the DTW matrix does not change and can be reused, because only the index of the matrix element where the final distance is read out changes.

gain, but the discriminativity-value, which is based on the k smallest distances to other subsequences of the training data (please refer to Eq. 6.1). The difficulty to utilize effective lower bounds on the DTW distances that actually save time has already been discussed in the previous section. This hinders an effective pruning of distance calculations: as the discriminativity computation does not use the distances to subsequences of *all* training videos, but the k *minimal* distances only, the discriminativity-value of a candidate under evaluation might improve notably by the consideration of so far unknown distances, as these distances might be better than the ones encountered so far and could thus replace the previous ones in discriminativity-value calculations, hence the discriminativity might improve significantly also very late in the computation process, depending on the order of evaluation. One possibility would be to use a threshold on the sum of the k minimal distances to subsequences of the opposite class (numerator in Eq. 6.1), as this value can only decrease during the computation process, thus reducing the discriminativity-value. However, this would be a heuristic which does not guarantee the optimal solution, because the denominator in Eq. 6.1 (the k minimal distances to subsequences of the same class) might also decrease, resulting in an increasing discriminativity-value, possibly outweighing the effect of a small numerator.

6.2.2. Logical Shaplets

Logical shapelets were developed by Mueen *et al.* [366] as an extension of the original shapelet approach [538]. They introduced conjunctions and disjunctions of shapelets which significantly increase the expressiveness when used for classification purposes. The new approach was evaluated on several datasets (hand signs in cricket, accelerometer data of a robot, graphical password trajectories, but no facial data) and found to outperform classical shapelets in these cases. In principle, this kind of logical combinations could also be incorporated in our dynamic recognition approach. However, it already comprises similar means of expressiveness. The training procedure selects t reference subsequences per class as prototypes, each of them can be matched to the test data at classification time (please refer to Sec. 6.1.2). This is equivalent to a disjunction of these prototypes. The classification decision is based on the distances to the u nearest prototypes (please see Sec. 6.1.3). This roughly corresponds to a conjunction of these prototypes, as the distances to all of them are considered simultaneously. The parameters t and u are automatically determined in the grid search during the model training.

Besides these logical combinations of shapelets, Mueen *et al.* [366] also presented two speedup techniques where they achieved in part notable performance enhancements. These techniques are discussed in the following two sections.

Efficient Distance Computation

Mueen *et al.* [366] presented a technique to speedup the distance calculations by intelligent caching and reuse of intermediate data. They suggested to compute some specific statistics for each time series of the training data that allow for Euclidean distance calculations in constant time later on [434]. The basic idea is to avoid repetitive computations of the same differences between two time series elements that arise from overlapping shapelet candidates.[8]

[8] Overlapping shapelet candidates are frequently encountered, because all subsequences of all training time series are possible shapelet candidates, so two "neighboring" candidates of the same length with slightly displaced starting points share most of their elements, thus the computed differences of these elements to the elements of other training times series can be reused.

In a sense, this method is similar to the integral image representation Viola and Jones [504] developed to quickly compute rectangular features for rapid object detection in images (please see Sec. 4.1.5).

However, this technique relies on a one-to-one relation between the elements of two time series that are compared during distance calculations as it is the case for the Euclidean distance. Naturally, such a relation does not hold for the DTW distance, which is why the DTW matrix needs to be recomputed whenever the starting point of a considered subsequence changes. Thus, this efficient distance computation technique is not applicable in our approach.

Candidate Pruning

Besides the efficient distance computation, Mueen *et al.* [366] also presented a new candidate pruning technique to further reduce the search space. The fundamental idea is that similar candidates—in terms of their Euclidean distance to each other—usually also lead to a similar information gain. Thus, if it is known that one candidate shapelet produces a very poor information gain, other candidates that are sufficiently similar to this one can be safely skipped. Formally, this "sufficient similarity" is analyzed exploiting the triangle inequality regarding the distance of the current candidate and a previously encountered "poor" candidate, finally resulting in an upper bound on the possible information gain of the current candidate. The speedup caused by this pruning method heavily depends on the data, ranging from none to a very significant one in the experiments reported by Mueen *et al.* [366].

Again, the utilization of this pruning method in our dynamic recognition approach is not unproblematic. First of all, the DTW distance does not fulfill the triangle inequality which is essential for this pruning technique. (However, it has been shown that for some applications in speech recognition on real data, the triangle inequality is at least loosely fulfilled in practice [502].) Moreover, as the discriminativity-value computation involves the evaluation of $2 \times k$ distances in total (Eq. 6.1) where the triangle inequality (if assumed to be approximately fulfilled) applies to each one individually and thus introduces a certain increase of the upper bound for each such distance which all sum up, the resulting upper bound on the discriminativity-value would not be as tight as the bound on the information gain where only one such distance needs to be considered. Hence this pruning method is expected to be less effective in our dynamic recognition approach.

Besides the aforementioned issues, some additional overhead would be unavoidable because our training approach also requires to process and hold *some* suboptimal candidates: due to the leave-one-out trial classifications during the parameter grid search (Alg. A.3), every suitable subsequence found—including the best candidate overall—is excluded once from the consideration, namely when the video it stems from is treated as unknown test data for parameter optimization. Of course, this is intrinsic to our approach and applies to the other pruning techniques discussed above as well.

6.2.3. Classification with Discriminative Subsequences by Nowozin *et al.* [380]

Nowozin *et al.* [380] also developed a classification approach based on discriminative subsequences. They use a part-based video representation where at each spatio-temporal voxel a spatial 2D Gabor filter and a temporal pair of 1D Gabor filters are combined to compute

spatio-temporal features [374]. These features are further processed to compute a descriptor of salient points, where the voxel values in a certain neighborhood around these points are concatenated in a large vector whose dimensionality is subsequently reduced via PCA. The resulting vectors are clustered via k-means to produce a codebook of prototypes [380, 375]. The indices in this codebook, together with the associated pixel coordinates and video frame numbers, form visual words as new representation. These words are temporally sorted and grouped into temporal bins, such that the final features used by the classifier become a sequence of sets of integers as representation of an input video [380].

The classifier is based on the *LPBoost* algorithm [111] where several weak hypothesis functions are linearly combined to form the classification function. This function is found by iteratively solving a linear programming problem which essentially involves maximizing a gain function in every iteration. To find the subsequence that maximizes this gain, Nowozin *et al.* [380] perform a tree search with a generalization of the *PrefixSpan* algorithm [398], embedded into a variation of the A^* search algorithm. This approach finds short subsequences with high gain early in the search process and uses this information to prune the search space, which is crucial to make the tree search tractable in practice [380]. The pruning is based on an upper bound on the maximum gain that might be achieved by further extensions of the currently investigated subsequence [364, 299]. This upper bound is tightly coupled to classifiers of the boosting family and is not easily applicable to different approaches, in particular not to our dynamic recognition approach that utilizes a nearest-neighbor-based classification with discriminative reference subsequences and uses dynamic time warping as distance measure. In fact, the approach of Nowozin *et al.* [380] and ours are entirely different, apart from the basic idea to use discriminative reference subsequences for classification.

The scenario Nowozin *et al.* [380] developed their classification approach for is human activity recognition. They presented very good results for the discrimination of boxing, handclapping, handwaving, jogging, running, and walking actions, evaluated on the KTH human action database [446]. To our knowledge, this approach has not been applied to face resp. FCS recognition, which also does not appear to be its primarily intended field of application due to the different demands: while FCS recognition requires the analysis of in part very subtle differences in the appearance of a very restricted object class at a comparatively high image resolution, the classification of human actions involves the analysis of much less restricted and probably less subtle motions at a relatively low resolution.

6.2.4. Facial Expression Recognition Approach of Buenaposada *et al.* [53]

Buenaposada *et al.* [53] also used a nearest-neighbor-based classifier (NN) to classify facial expressions. They organized the face tracking in an three-layer architecture: detection of skin-colored blobs [51], verification whether these blobs are faces by a template-based rigid face tracker [52], and detailed face modeling by a subspace-based tracker that handles both facial expressions and illumination changes by means of two independent linear models. The subspace-based tracking is done by minimizing an error function to fit the model to the data using the Gauss-Newton algorithm [35], where a suitable factorization of the Jacobian matrix enables online performance by precomputing large parts of the required quantities offline in advance. They applied this tracker to video sequences from the Cohn-Kanade database [266] showing six basic emotions [133] to build models of facial expressions, where each emotion category is represented by a trajectory in the space of deformation parameters. Subsequently,

they apply linear discriminant analysis (LDA) [127] to reduce the dimensionality of these trajectories.

The classification of test data is realized by a k-NN classifier that is used to estimate the probability of the facial expression categories, given the test data and the facial expression models computed during training. The prototypes are points of the trajectories in the linear facial expression subspace which correspond to a single face image each. The temporal dynamics are considered during the computation of the posterior probabilities where the respective probabilities of preceding frames are taken into account, in contrast to our approach where the DTW distance accounts for these dynamics. Despite the similarities of modeling the facial expressions with linear deformation models and the usage of a k-NN classifier, the overall approach of Buenaposada *et al.* [53] is rather different from ours regarding most other aspects, including the prototype selection method.

6.2.5. Conclusion

In this review of related work, we considered several classification methods that are based on discriminative subsequences or utilize an k-NN classifier. The discussion focused on the differences to our approach and the difficulties in equipping it with effective pruning strategies similar to those used in these related methods. Although in our dynamic recognition approach the training is done offline anyway, further speedup techniques in addition to the optimizations explained in Sec. 6.1.6 are desirable for its practical usage and are thus subject to future work.

6.3. Evaluation

This section evaluates the developed dynamic recognition approach on the videos of the object-teaching database (please see Sec. 3.2). We used the same face detection and AAM feature extraction methods as in the static baseline approach (please refer to Sec. 5.1 and Sec. 5.2.1). First of all, we investigate the person-specific FCS classification with individual AAMs in detail in Sec. 6.3.1. Subsequently, the usage of generic AAMs is evaluated in Sec. 6.3.2, before the generalization to new persons is considered in some tentative experiments in Sec. 6.3.3. Finally, Sec. 6.3.4 presents the results of additional experiments where the AAMs were replaced by CLMs (please refer to Sec. 4.5.2). Please refer to App.A.7 for a list of the dimensionalities of the feature vectors that were used in these classification experiments.

6.3.1. Person-Specific Classification with Individual AAMs

This section reports our experiments regarding the person-specific classification with individual AAMs. For each person, an individual AAM was trained, based on approximately 200 face images of this person with hand-annotated feature points, likewise to the experiments reported in Sec. 5.2.1 and Sec. 5.4. As before, the AAM parameter vector gained from the fitting of this model to the face was used as feature vector for the respective frame. Due to the dynamic nature of the recognition approach where sequences of feature vectors are compared instead of single frames, missing frames can hinder the classification more than in the static approach, because a certain minimal length of consecutive frames is required to form a meaningful subsequence, whereas this is no problem for the static approach reported in chapter

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	73	86	83	91	72	50	75	82	54	71	93	75.4	13.8
AAM-96 – success	67	88	88	94	63	43	67	93	42	58	88	71.8	19.5
– failure	78	83	75	88	81	58	81	68	61	83	97	77.6	11.5
– all	94	83	82	88	83	89	89	65	67	74	88	81.9	9.4
AAM-99 – success	86	82	95	88	82	90	96	58	37	83	87	80.2	17.7
– failure	100	83	64	88	85	88	84	73	83	64	89	81.7	10.9
– all	93	95	92	100	92	80	87	92	83	97	99	91.8	6.3
T-96 – success	90	95	96	100	87	78	89	92	64	96	97	89.5	10.4
– failure	95	94	86	100	96	82	86	92	93	98	100	92.9	6.0
– all	97	94	83	100	85	77	92	87	88	92	97	90.2	6.9
T-99 – success	93	93	100	100	100	100	94	92	88	97	99	96.0	4.2
– failure	100	97	57	100	70	51	90	81	88	86	95	83.2	17.0
– all	94	86	83	88	72	50	89	82	67	71	93	79.5	13.3
AAM-S – success	86	88	88	88	63	43	96	93	37	58	88	75.5	21.1
– failure	100	83	75	88	81	58	84	68	83	83	97	81.8	11.9

Table 6.3.: Classification accuracies for AAMs with 96% and 99% variance preservation in a person-specific classification with individual AAMs (rows *AAM-96* and *AAM-99*). Additionally, the mean leave-one-out training accuracies of the two AAM variants are shown (rows *T-96* and *T-99*). The last row (*AAM-S*) shows the results when one of the two AAM variants is selected for a person based on the achieved training error. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 6.3.1.

5. To cope with this issue, we closed short gaps in the frame sequences by linearly interpolating between the surrounding feature vectors. This makeshift enables the consideration of subsequences that would otherwise be rejected due to missing frames.

We first consider the classification in a leave-one-out cross validation manner, before we discuss the stability of the involved classifier parameters. Afterwards, the achieved results are compared to those of the static baseline approach (please see Sec. 5.4) and the human recognition performance (please see Sec. 3.5.2). Subsequently, the reference subsequences selected in the classifier training are evaluated.

Leave-One-Out Cross Validation over Scenes

We performed a leave-one-out classification of all scenes for each person separately, i.e. each scene was treated as test data once, where the classifier was trained with all remaining scenes. The resulting classification accuracies are shown in Tab. 6.3. Two variants of the AAMs were used: one that preserves 96% of the variance of the training data (*AAM-96*), and one with 99% variance preservation (*AAM-99*). The latter achieved a good average classification accuracy of 81.9%, which is considerably better than the best results of the static approach and comparable to the human performance. Interestingly, despite performing better on average, the *AAM-99* variant did not consistently outperform the *AAM-96* variant for all persons, quite the contrary is the case: the former performed better for six subjects, the latter for the remaining five. Consequently, there is no significant correlation at all between the results for

person ↓ | median para. over scenes →	$(l_{\min}, l_{\max})$	k	t	v	u	w	b
01	(5,5)	1	2	0	1	1	1.1316
02	(5,20)	5	5	0	4	1	0.8990
03	(10,10)	1	25	1	2	2	0.8291
04	(5,5)	1	15	0	5	1	0.8162
05	(5,5)	1	5	0	1	1	0.7822
06	(5,5)	5	5	0	2	1	0.9351
07	(10,10)	15	15	0	6	1	1.0558
08	(5,5)	5	10	1	3	1	0.8783
09	(5,5)	15	5	0	1	1	1.4368
10	(5,5)	1	25	2	2	1	1.0426
11	(5,5)	15	20	2	7	1	0.9811
median parameters over persons	(5,5)	5	10	0	2	1	0.9663

Table 6.4.: Median classifier parameters for all scenes of each person (rows *01–11*) and the median parameters over all persons (last row). Please refer to Sec. 6.3.1 and Tab. 6.1.

the individual persons of these two AAM variants (Spearman correlation, $\rho \approx 0.06$, $p > 0.86$). The reason for these somewhat surprising differences is not yet understood.

We also evaluated the mean leave-one-out training error, i.e. each training video was treated as unknown test data once and the classifier was trained on the remaining training videos only. The mean classification accuracies resulting from this procedure are shown in the rows *T-96* and *T-99* of Tab. 6.3. The question we were interested in is whether this training error can be used to assess the quality of the trained classifier and thus to estimate its expected performance on test data. It turned out that there is a significant correlation between the training and test performance for the *AAM-96* features (Spearman correlation, $\rho \approx 0.70$, $p < 0.02$). Unfortunately, this does not hold for the *AAM-99* variant (Spearman correlation, $\rho \approx 0.33$, $p > 0.3$).[9] However, the main reason of this lack of significance is the performance for subject 06, the only case were the training error is worse than the test error. For the ten other subjects, the correlation between training and test classification accuracy is significant (Spearman correlation, $\rho \approx 0.64$, $p < 0.05$). So far, the reason for the atypical situation with subject 06 is unclear. Apparently, the leave-one-out training error can help to estimate the quality of the trained classifier in terms of expected performance on test data for most persons,[10] but there are exceptions that are to be investigated in future work.

One simple improvement idea is to select the AAM variant to use for the respective person based on the better training error. However, this did not yield better test classification accuracies on average, as row *AAM-S* in Tab. 6.3 shows,[11] because the "wrong" decision was

[9] This correlation is significant for the *success* scenes only: $\rho \approx 0.62$, $p < 0.05$

[10] A simple heuristic for this estimation would be to use a threshold on the leave-one-out training accuracy of 90%. This leads to a mean test classification accuracy of 86.0% for the six people above the threshold and a mean test classification accuracy of 77.2% for the five people below it in case of the *AAM-99* features. For the *AAM-96* variant, this results in 81.4% for the eight people above the threshold and 59.7% for the three people below it. However, more sophisticated ways to assess the expected test performance are to be investigated in future work.

[11] In case of equal training errors (as it was the case for subject 04), the *AAM-99* features were chosen due to their better average performance.

median para.	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	97	83	91	91	92	83	89	71	82	87	90	86.8	6.9
scenes – success	93	83	95	100	100	80	96	62	64	83	87	85.6	13.3
– failure	100	83	86	81	85	88	84	82	91	91	91	87.4	5.6
– all	75	69	74	84	71	78	71	75	56	65	78	72.3	7.5
persons – success	93	59	75	69	45	100	71	65	100	100	83	78.2	18.5
– failure	61	83	71	100	92	50	71	86	35	27	74	68.4	23.2

Table 6.5.: Classification accuracies for two special parameter selections: the median classifier parameters for all scenes of a person (row *scenes*), and again the median parameters of these median parameters (row *persons*). For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 6.3.1.

made in some cases, rendering this simple scheme insufficient. Nevertheless, the basic idea of person-specific feature extraction model selection is promising and subject to future research. In the subsequent evaluations, we focus on AAMs that preserve 99% variance only, because of their better performance on average.

Classifier Parameter Stability

The leave-one-out cross validation classification experiments yielded one set of parameters for each scene. As already argued in the evaluation of the static approach, a certain parameter stability is necessary for the practical usage of a classification system, because a classifier trained with a particular parameter set needs to yield good results on various test data. To estimate this stability, we computed a single parameter set for all scenes of a person by using the median value[12] of the single cross validation parameter sets for each parameter, except for the classification bias b where the geometric mean was used instead of the median. The reasoning behind this is that, if a sufficient stability is present, the slightly different training data sets in the leave-one-out cross validation classifications of the single scenes should yield slightly different parameter sets, which on average capture some characteristics of the respective person. Thus, taking the median value of each parameter should be a good guess for a single parameter set that yield good results for all scenes.

Table 6.4 lists the resulting median parameter sets, and the respective classification accuracies are shown in the row *scenes* of Tab. 6.5. Compared to the cross validation results, the classification accuracies improved for most of the persons. However, these numbers are not meant to be taken for the evaluation of the classifier in terms of absolute classification accuracy. As the median operation is performed on the parameter sets of *all* scenes, it also processes information extracted from the respective test data, which is a likely reason for the performance improvement (please see the respective experiments for the static approach in Sec. 5.4.1). The point here is that a single parameter set with plausible values (median

[12] More precisely, the median was used if the number of scenes was odd. For an even number of scenes, not the mean value of the two middle elements was computed, but the larger of these elements was used directly as "median" value of the given parameter set. This was done to ensure that the resulting median value is actually an element of the input set, because apart from the classification bias b, all parameters of the classifier are integer, thus using the mean value of the two middle values might yield a non-integer value which is not applicable as parameter.

approach/features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	76	83	80	95	84	57	62	74	66	71	88	76.0	11.5
static – success	67	82	89	90	81	60	52	69	25	75	83	70.3	19.2
– failure	83	83	67	100	88	54	70	81	87	67	91	79.2	13.3
– all	73	86	83	91	72	50	75	82	54	71	93	75.4	13.8
AAM-96 – success	67	88	88	94	63	43	67	93	42	58	88	71.8	19.5
– failure	78	83	75	88	81	58	81	68	61	83	97	77.6	11.5
– all	94	83	82	88	83	89	89	65	67	74	88	81.9	9.4
AAM-99 – success	86	82	95	88	82	90	96	58	37	83	87	80.2	17.7
– failure	100	83	64	88	85	88	84	73	83	64	89	81.7	10.9
– all	82	75	85	92	68	73	94	67	78	95	92	82.0	19.1
human – success	91	66	84	89	61	70	91	52	66	95	93	78.1	21.2
– failure	73	84	86	95	75	75	98	82	91	95	91	86.0	16.1

Table 6.6.: Classification accuracies of the static recognition approach using the mean AAM features, the dynamic recognition approach using the *AAM-96* features, the dynamic recognition approach using the *AAM-99* features, and the human recognition performance. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 6.3.1.

values, see argumentation above) yielded a reasonable good performance for all scenes of a person. This is an indication that stable parameters exist for each person. For the purpose of evaluating the actual performance of the classifier, the cross validation results discussed above are determinative.

An interesting question is whether a single stable parameter set can also be selected for all persons. The partially large differences between the median parameter sets for different persons (Tab. 6.4) let us doubt this. This negative expectation is confirmed by a tentative experiment where we computed again the median values of all the median parameter sets, resulting in a single parameter set for all persons. This parameter selection impairs the classification results notably, as the row *persons* in Tab. 6.5 shows. Thus, suitable parameters of the classifier appear to be person-specific and do not generalize well to other persons.

Comparison to the Results of the Static Baseline Approach

This section compares the classification accuracies achieved by the dynamic approach to the best results of the static approach, namely the classification results yielded by the mean AAM feature vectors (please refer to Sec. 5.4.2). For convenience, these results are shown again in Tab. 6.6. Using the *AAM-99* features, the dynamic approach outperformed the static one on average (81.9% vs. 76.0% for all scenes, 80.2% vs. 70.3% for *success* scenes, and 81.7% vs. 79.2% for *failure* scenes) and for six people. For two more persons, both performed equally well, and for the remaining three people, the static one yielded better results. Due to these differences and the high variance in general, the better performance of the dynamic approach is not statistically significant ($p > 0.19$ for both a two-tailed t-test and a Wilcoxon rank sum test). However, we suppose that the better average performance is real, but the significance cannot be confirmed due to this high variance for different persons. A substantial larger number of subjects would be necessary to test this hypothesis.

A closer comparison of the classification accuracies of the static and the dynamic approach for both *AAM-96* and *AAM-99* features leads to a surprising discovery. We already noted above that the *AAM-96* and *AAM-99* features yielded quite different results for the individual people. The classification accuracies for different persons of the dynamic approach using *AAM-99* features are also very different from those of the static one (Spearman correlation, $\rho \approx 0.005$, $p > 0.98$). In sharp contrast, the classification accuracies of the dynamic approach using *AAM-96* features and the ones of the static approach are significantly correlated (Spearman correlation, $\rho \approx 0.76$, $p < 0.01$). In general, the dynamic approach with the *AAM-96* features performed very much like the static one with the mean AAM features:

- the average classification accuracies for all scenes, *success* scenes, and *failure* scenes are very comparable (just 0.6 – 1.6 percentage points difference)
- the classification accuracy is considerably better for the *failure* class than for the *success* class
- the variance of the classification accuracy for different subjects is considerably higher for the *success* class than for the *failure* class
- the classification accuracies for the individual subjects are significantly correlated

On the contrary, the dynamic approach with the *AAM-99* features compares very different to the static one: only the classification accuracy for the *failure* class is comparable, the difference for the *success* class is 9.9 percentage points (thus the dynamic approach performed almost equally well for *success* and *failure* in this case), and the results for individual persons do not appear to be correlated. Only the higher variance for *success* than *failure* scenes remains as commonality. The reason for this apparent switch in the behavior of the dynamic recognition approach with increased feature vector dimensionality is not yet understood and subject to future research. A particular question to investigate is for which kinds and distributions of facial displays this behavior occurs—and why.

Comparison to the Human Recognition Performance

The average classification accuracy of 81.9% of the dynamic recognition approach using *AAM-99* features is comparable to the average human performance of 82.0%, the associated classification accuracies are listed in Tab. 6.6 again, for convenience. There is no significant correlation between the respective classification accuracies of the individual persons (Spearman correlation, $\rho \approx 0.24$, $p > 0.47$), albeit the correlation might be stronger than those of the static approach and the human performance (please see Sec. 5.4.3). When the performance of the *AAM-99* features is evaluated on those 88 videos only which the human performance is based on (please see Sec. 3.5.1), the results are very similar: 81.1% for all videos (SD 15.7), 84.1% for *success* videos (SD 20.2), and 79.6% for *failure* videos (SD 24.5). This supports our suspicion that the increased performance of the static approach on this subset is by chance, not because the humans judged an in a sense "easy" subset of the videos. Likewise to the comparison of the human recognition performance and the best results of the static approach, we also evaluated the correlation of the classification results on the level of the single videos instead on the level of average performance per person (please see Sec. 5.4.3). It turned out that there is a weak, but significant correlation between these classification results on this subset of videos (Spearman correlation, $\rho \approx 0.29$, $p < 0.01$). Thus, measured on video level, the human observers and the dynamic recognition approach tended to make some similar classification errors to some (weak) extent (similar to the situation with the static approach).

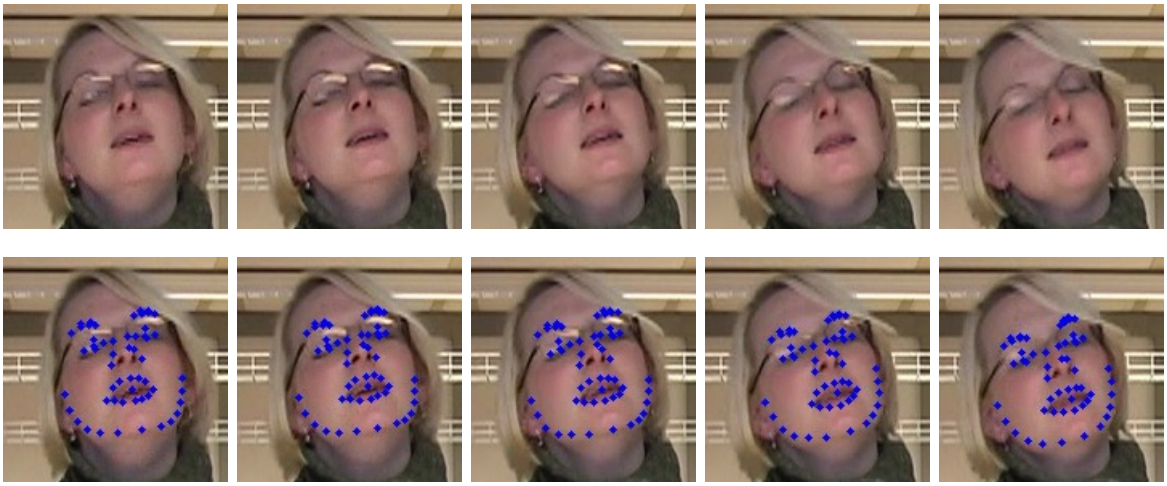
Evaluation of Reference Subsequences

We visually inspected the reference subsequences that were selected by the classifier training with the *AAM-99* features and compared them to the facial displays of the respective person that were evaluated in Sec. 3.4. As example illustrations, the best reference subsequences for *success* and *failure* for five subjects of the object-teaching study are depicted in Fig. 6.4 to Fig. 6.8. The "best" reference subsequences are those with the highest discriminativity-value (please refer to Sec. 6.1.1). All depicted reference subsequences happened to be five frames long, so the figures show all of their frames.

It turned out that the reference subsequences indeed capture behaviors that are typical for the person (please refer to the captions of Fig. 6.4 to Fig. 6.8). On the one hand, this does not come at a surprise, as the reference subsequences by definition exploit characteristic differences between *success* and *failure* that are typical for a particular person. On the other hand, there are usually several behaviors which are typical for a person, from the analysis conducted in Sec. 3.4 one cannot easily tell which will be the most discriminative ones that will be selected by the training procedure.

The reference subsequences comprise head gestures, in particular head nods (e.g. subjects 06 and 10), gazing behavior, especially gazing at the robot, at the object, or downwards to the table (e.g. subjects 06, 08, and 11), and facial expressions, mainly normal vs. pronounced speech (e.g. subjects 08, 09, and 11). Some of these behaviors appear to be specific to the investigated scenario resp. context (please refer to Sec. 3.7). For instance, looking down to the table with objects after a successful interaction, in order to put the object back and possibly choose a new one, is probably tightly coupled to this object-teaching scenario or very similar situations. In contrast, looking at the interaction partner and using pronounced speech in the presence of a communication problem is presumably a typical behavior for a wider range of scenarios. As argued in Sec. 3.7, the investigation of further scenarios is required to conclusively assess the specificity or generality of the shown FCSs.

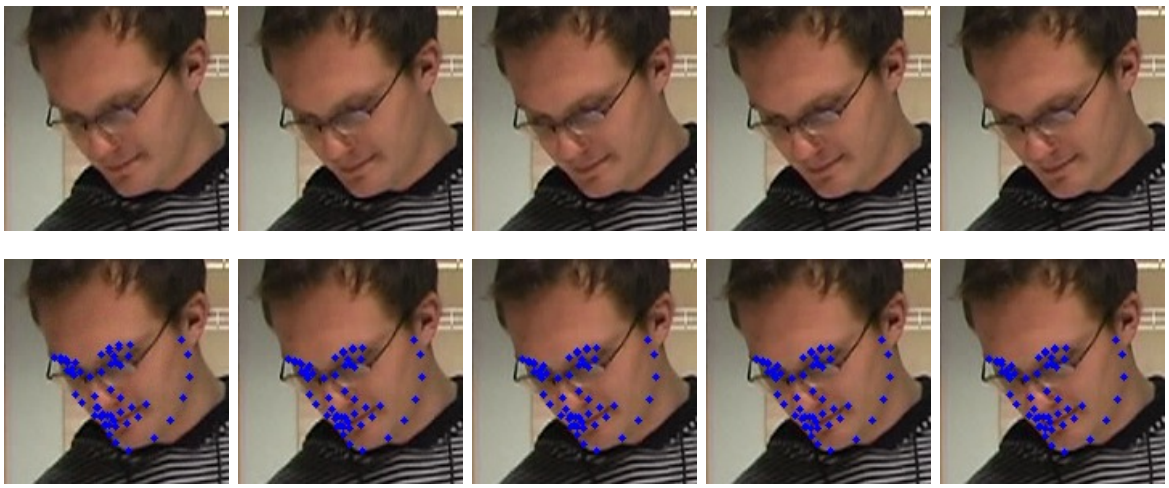
We did not find a simple relation between the kind of facial displays shown in the reference subsequences and the achieved classification accuracy. Very similar reference subsequences of the *success* and *failure* classes are likely to yield relatively poor classification accuracies, due to the high risk of confusion. As the reference subsequences are chosen to be very discriminative, this reflects the difficulty of the classification problem for the respective person. This is the case for subjects 09 and 10. Their reference subsequences for *success* and *failure* are rather similar, consequently the comparatively poor classification results for these persons are not surprising. In contrast, very different reference subsequences for the two classes do not to the same degree allow to expect a good classification performance, because there are several other causes for poor results: the best reference subsequences might not be matched very often, so that the classification needs to consider less discriminative ones, or the AAM fitting might yield too poor results in too many cases, for instance. Two contrasting examples are the subjects 08 and 11. In both cases, the reference subsequences for *success* and *failure* are rather different. Furthermore, the reference subsequences of the same class appear to be similar for these two people. Nevertheless, the classification accuracy for subject 08 is poor (65%), but the classification accuracy for subject 11 is good (88%). A closer investigation of this issue, for instance by analyzing in detail which subsequences are matched how often and what is the result of this regarding classification scores and accuracy, is subject to future work.



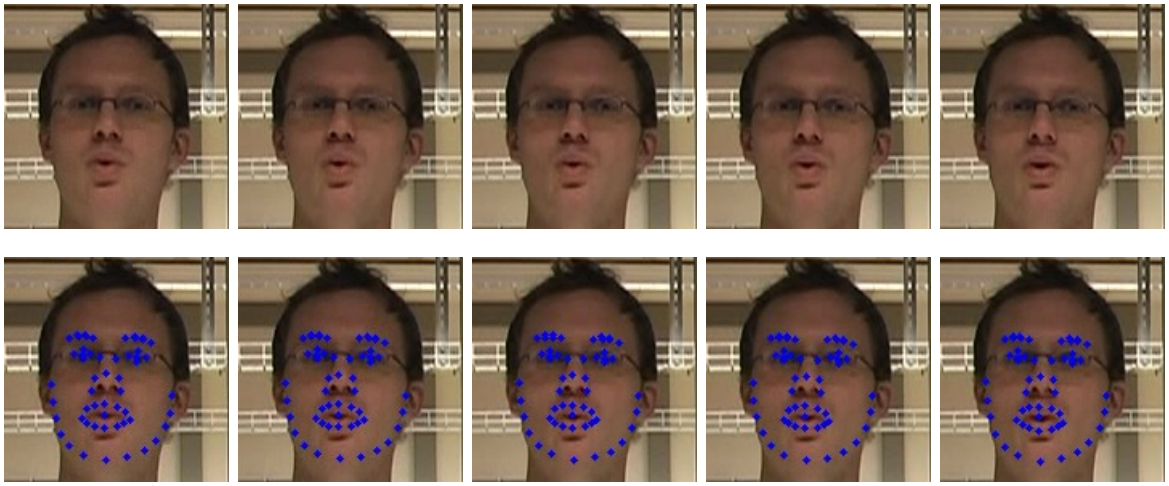
Best reference subsequence for *success* scenes

Best reference subsequence for *failure* scenes

Figure 6.4.: Best reference subsequences for subject 06. Top: Best reference subsequence for *success* scenes. This is the beginning of a nodding head gesture, overlayed with a "roll" movement of the head which is typical for this person. Bottom: Best reference subsequence for *failure* scenes. The person is gazing towards the object she holds in her hands. Both behaviors are typical for this person, as she used head nods in about three out of four *success* scenes and also gazed towards the object in almost half of the *failure* scenes (please see the evaluation in Sec. 3.4). Please refer to Sec. 6.3.1.

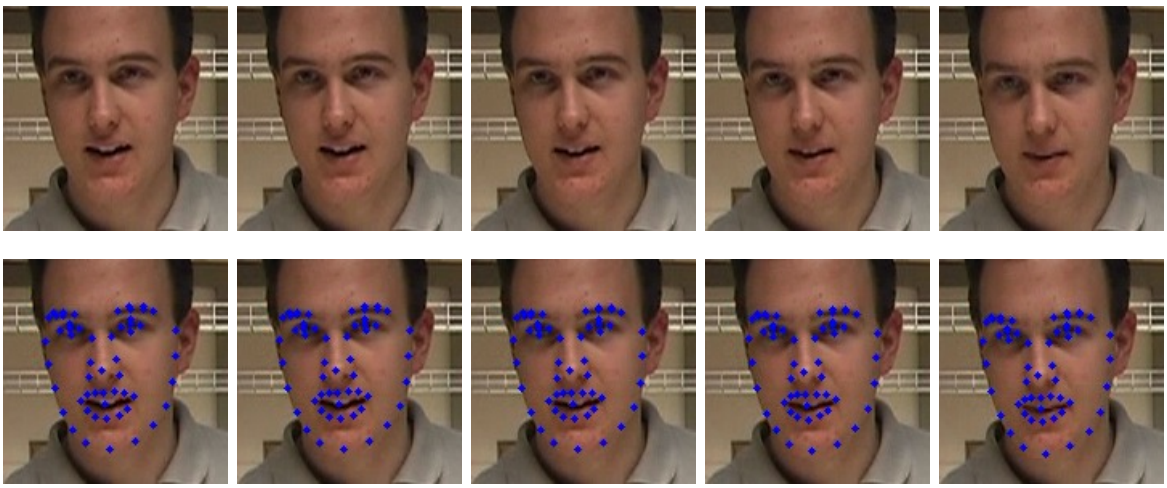


Best reference subsequence for *success* scenes

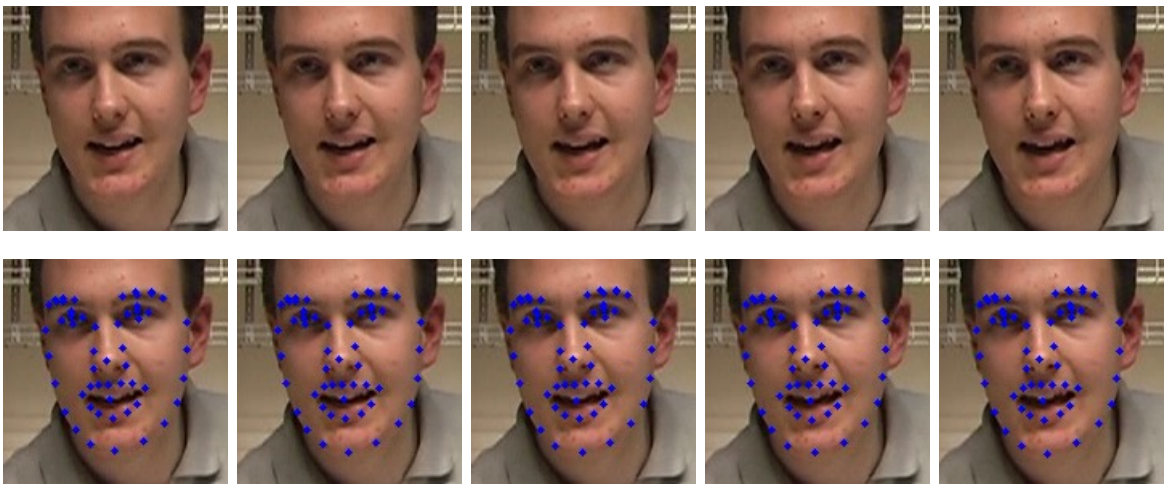


Best reference subsequence for *failure* scenes

Figure 6.5.: Best reference subsequences for subject 08. Top: Best reference subsequence for *success* scenes. The person looks downwards to the table with objects to put the object back. Bottom: Best reference subsequence for *failure* scenes. The person looks at the robot and corrects the wrong answer of the robot by uttering the correct object name in a pronounced way. Both behaviors are typical for this person, as he gazed downwards to the object table in about three out of four *success* scenes and also used pronounced speech while gazing at the robot in about three out of four *failure* scenes (please see the evaluation in Sec. 3.4). Please refer to Sec. 6.3.1.

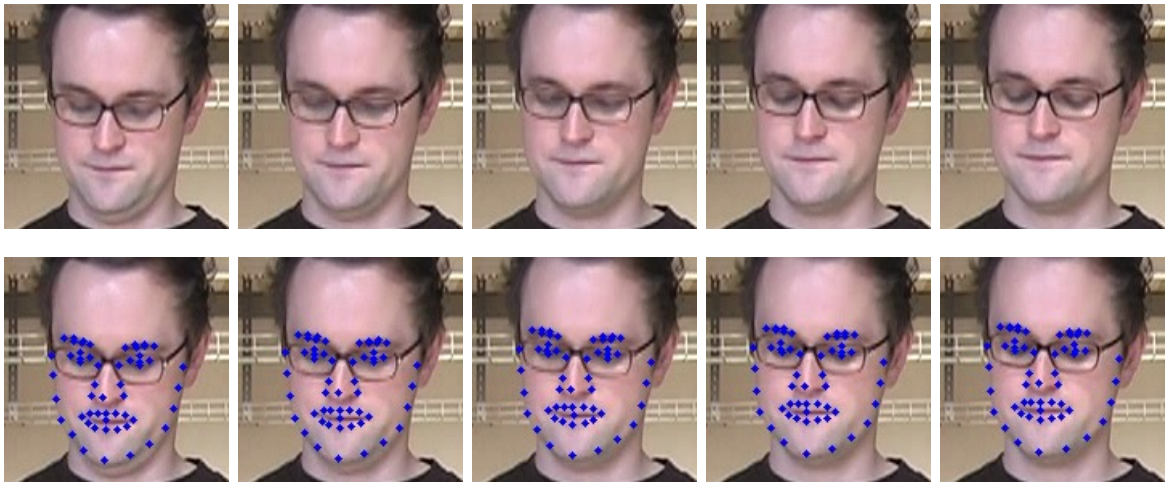


Best reference subsequence for *success* scenes

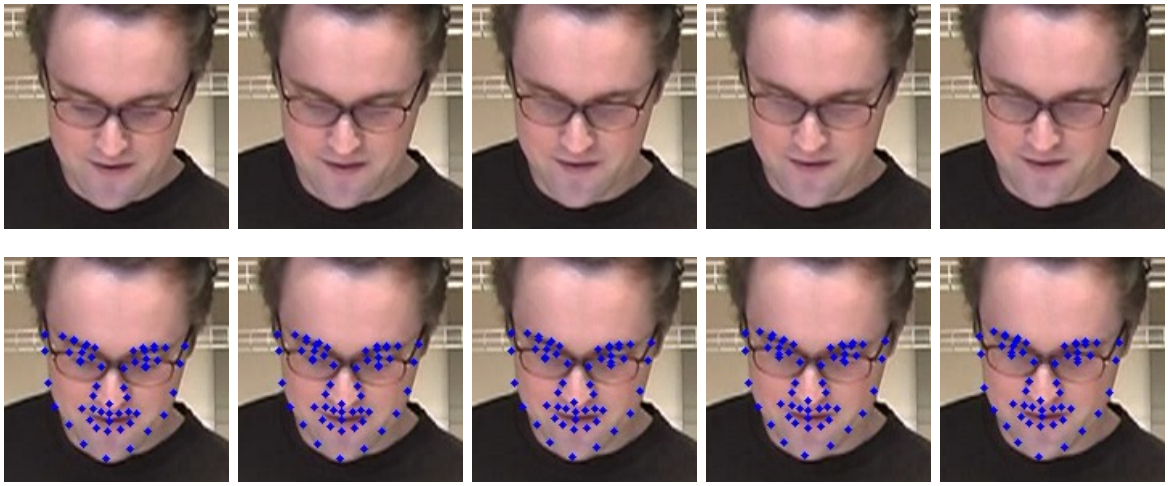


Best reference subsequence for *failure* scenes

Figure 6.6.: Best reference subsequences for subject 09. Top: Best reference subsequence for *success* scenes. The person looks at the robot and confirms the correct answer using normal speech. Bottom: Best reference subsequence for *failure* scenes. The person looks at the robot and corrects the wrong answer of the robot using a pronounced speech. Both behaviors are typical for this person, as he gazed to the robot in all *failure* scenes and 92% of the *success* scenes, and never used pronounced speech in a *success* scene, but in 92% of the *failure* scenes (please see the evaluation in Sec. 3.4). Please refer to Sec. 6.3.1.

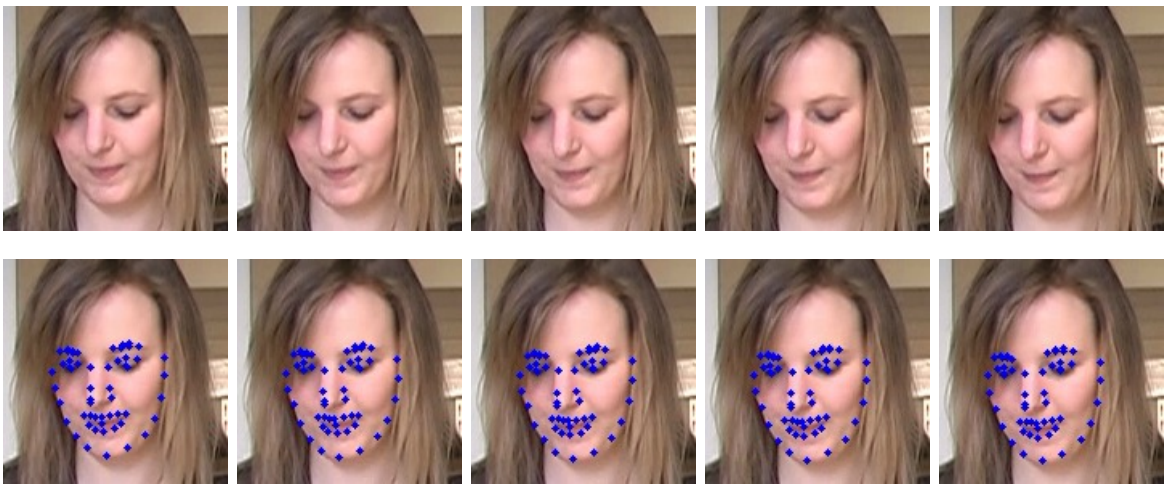


Best reference subsequence for *success* scenes

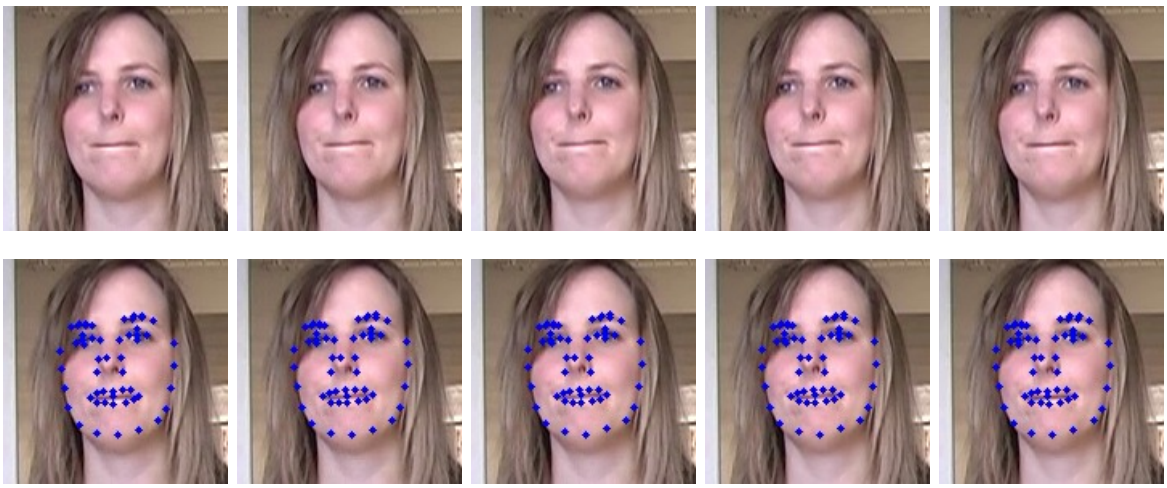


Best reference subsequence for *failure* scenes

Figure 6.7.: Best reference subsequences for subject 10. Top: Best reference subsequence for *success* scenes. This subsequence is part of a nodding head gesture. Bottom: Best reference subsequence for *failure* scenes. This subsequence is part of a slight upward movement of the head which this person often performed before he verbally reacted to the wrong answer of the robot. Both behaviors are typical for this person, as he nodded in 83% of the *success* scenes, but also used upward or downward head movements in 92% resp. 83% of the *failure* scenes (please see the evaluation in Sec. 3.4). Please refer to Sec. 6.3.1.



Best reference subsequence for *success* scenes



Best reference subsequence for *failure* scenes

Figure 6.8.: Best reference subsequences for subject 11. Top: Best reference subsequence for *success* scenes. The person looks downwards to the table with objects to put the object back. Bottom: Best reference subsequence for *failure* scenes. The person looks at the robot and begins to correct the wrong answer of the robot using pronounced speech. Both behaviors are typical for this person, as she gazed downwards to the object table in 88% of the *success* scenes and also used pronounced speech while gazing at the robot in 94% of the *failure* scenes (please see the evaluation in Sec. 3.4). Please refer to Sec. 6.3.1.

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	79	79	70	91	66	52	76	78	66	58	85	72.7	11.6
G-AAM – success	87	76	83	88	63	50	75	87	33	50	75	69.8	18.3
– failure	72	83	50	94	69	54	77	68	63	67	91	73.5	14.0
– all	73	90	68	73	63	46	80	76	54	83	68	70.3	12.7
L-AAM – success	80	94	75	76	63	43	83	80	42	75	58	69.9	16.7
– failure	67	83	56	69	63	50	77	72	61	92	75	69.4	12.1

Table 6.7.: Classification accuracies in a person-specific classification with generic AAMs. Row *G-AAM* shows the results for an AAM that contains images of all 11 people, while row *L-AAM* lists the results for AAMs that leave out the target person, thus comprising images of the 10 remaining people. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 6.3.2.

6.3.2. Person-Specific Classification with Generic AAMs

In all experiments so far, we used person-specific, individual AAMs, because they are known to yield better fitting results than generic models in general [198]. However, in case one wants to recognize the FCSs of a new, yet unknown person, generic models are required. Therefore, we built one generic AAM (with 99% variance preservation) from the annotated images of all 11 people in the object-teaching database to test its performance. The classification accuracies resulting from this model are shown in row *G-AAM* of Tab. 6.7. They are notably impaired by the feature extraction of the generic AAM, the average accuracy of 72.7% is 9.2 percentage points below the average accuracy of the individual models. As this generic AAM still contains images of the respective target person, we built additional models were the images of the target person are left out. The resulting classification accuracies are shown in row *L-AAM* of Tab. 6.7. The average accuracy decreased once more to now 70.3%.

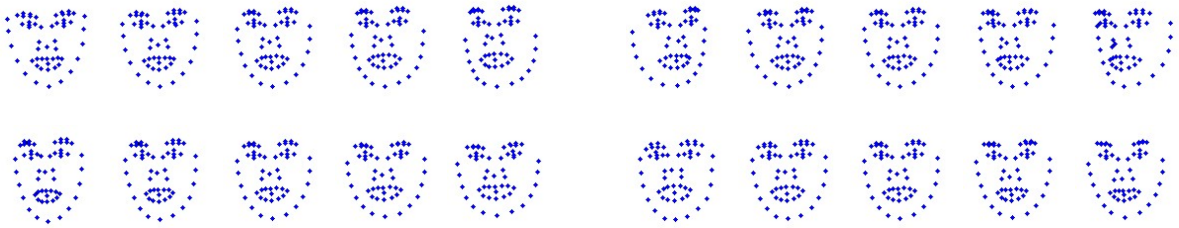
Unfortunately, the difference between the individual models and the generic model for all persons is much bigger than the difference between this generic one and those that leave out the target person. It seems that a large amount of the variance captured by the generic AAM accounts for the "other" people in the training data, thus reducing the represented variance for the respective target person. To some degree, this can be seen from the illustration of the generic AAM in Fig. 6.9. Two of the first four parameters are tightly related to the identity of the modeled persons, whereas variations in eye gaze and facial expression are less precisely represented, compared to the individual AAM shown in Fig. 5.3. To compensate for such an effect, a higher variance preservation can be chosen, which, however, results in feature vectors of considerably higher dimensionality, thus more training data might be required for a successful training of a classifier of good quality.

6.3.3. Generalization to New Persons

In all experiments so far, we considered a person-specific classification only. The generalization of a trained classification system for FCS recognition to new, unknown people is very desirable, but also very challenging. Based on the results of the preceding evaluations, we doubt that a classifier trained with the dynamic recognition approach will generalize well to other persons in object-teaching database:

Variation of the appearance parameters (combined shape and texture parameters)

Variation of the texture parameters



Variation of the shape parameters

Figure 6.9.: Example illustration of the generic AAM comprising the annotated images of all 11 subjects of the object-teaching study. The upper block shows the model variance represented by the combined appearance parameters of shape and texture, the middle block the variance represented by the texture parameters, and the lower block the variance represented by the shape parameters. In each block, the first four parameters are varied and the effect is shown in a series of five images in each case: $\bar{x} - 2\sigma$, $\bar{x} - \sigma$, $\bar{x}$, $\bar{x} + \sigma$, and $\bar{x} + 2\sigma$ (from left to right), where $\bar{x}$ is the mean value of the respective parameter and σ its standard deviation. The top left series depicts the variation of the first parameter, the top right the second ones, the bottom left the third ones, and the bottom right the fourth ones. The parameters are sorted according to the variance of the model training data they represent. It can be seen that the first two shape parameters mainly represent head poses, while the following ones correspond to shape variations induced by the identities of the modeled people. The first two texture parameters mainly represent slight changes in illumination, the third and fourth parameter different identities and slight mouth movements. Thus, the first four combined appearance parameters of shape and texture represent head movements, identity, and mouth movements as dominating facial expressions. Please refer to Sec. 6.3.2.

- The participants of the object-teaching study showed a large variety in their displayed FCSs (please refer to Sec. 3.4)
- Well-suited classifier parameters turned out to be rather diverse for different people (please refer to Sec. 6.3.1)
- The selected reference subsequences actually reflect the differences between the individual persons[13] (please refer to Sec. 6.3.1)
- Already the usage of generic AAMs in a person-specific classification notably impaired the classification results, thus even worse results can be expected for a person-independent classification (please refer to Sec.6.3.2)

Nevertheless, we conducted some tentative experiments regarding the generalization to new persons for the sake of completeness. We performed a leave-one-out cross validation over the 11 people in the object-teaching database, where all scenes of one person were treated as test data, while the classifier was trained using all scenes of the remaining persons. The most straightforward way to conduct the training would be to join all scenes of all training persons into one large training set. Unfortunately, our prototype implementation in Matlab would require too much working memory to handle training datasets this large, a particular issue is the precomputation and usage of a large number of minimal subsequence distances, where a lot of auxiliary data structures are used by Matlab (please refer to Sec. 6.1.6 and Alg. A.6). The memory requirements could be significantly reduced by re-implementing the dynamic recognition approach in a programming language with a more explicit control of the memory management, such as C++. Furthermore, it is also possible to hold only considerably smaller parts of the required data structures in the working memory and load the needed data on demand, this can be efficiently incorporated in the training process (Alg. A.1, A.3, A.6, and A.8 would be adapted). However, such an implementation was not available at the time of this writing.

Thus, we performed the training in another way. An individual classifier for each person was trained as before, using all videos of the respective person. The videos of the test person were classified by a combination of these classifiers by simple heuristics, as explained below. We used the generic *G-AAM* variant in these experiments, because it dramatically reduced the training effort compared to the *L-AAM* variant which leaves out the target person. In the former case, the same classifiers can be used for all classifications, just the classifier trained on videos of the test person is left out. In the latter case, for each new test person, the classifiers of all training persons need to be retrained, because they now operate on different features, namely the parameter vectors of a new AAM that leaves out the target person. This results in a squaring of the training time. However, according to the results reported in Sec. 6.3.2, slightly lower classification accuracies are to be expected from the usage of the *L-AAM* variant.

Fusion of Classification Results

The fusion of classification results is done by a simple majority voting over the classification results of the individual classifiers trained on one training person each. These classifiers serve

[13] The best reference subsequences selected for subject 10 and subject 08 or 11 are an extreme example for these large differences: the best reference subsequence for *failure* of subject 10 appears to be very similar to the best reference subsequences for *success* of subjects 08 and 11, and also the best reference subsequence for *success* of subject 10 resembles the best reference subsequences for *failure* of subjects 08 and 11 (please refer to Fig. 6.5, 6.7, and 6.8).

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	64	45	45	76	44	41	65	51	57	50	63	54.5	11.2
G-AAM-C – success	40	53	42	59	31	79	50	27	75	58	71	53.1	17.3
– failure	83	33	50	94	56	0	77	80	48	42	57	56.4	26.8
– all	61	41	70	85	69	41	71	64	71	54	58	62.2	13.3
G-AAM-S – success	33	24	79	88	88	79	58	47	67	42	79	62.1	22.8
– failure	83	67	56	81	50	0	81	84	74	67	43	62.3	25.0

Table 6.8.: Classification accuracies for a person-independent classification with generic AAMs. Row *G-AAM-C* lists the results for a fusion of classification results and row *G-AAM-S* for a fusion of classification scores. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 6.3.3.

as a model of the discriminative characteristics of the respective person. The test scene is first classified with all these models and finally classified into the most frequently occurring class of these classifications (ties were counted as wrong classifications in the conducted experiments).

The reasoning and hope behind this simple method is the following. Most people in the training set are probably rather different from the test person. Some of the associated models might classify scenes of the test person at random, some might show a strong bias towards *failure*, others to *success*; in any case the classification results of these models are impractical. However, on average, their classification results might cancel out each other. In case there are also a few people in the training set who are rather similar to the test person, the associated models might classify scenes of the test person reasonably well, at least significantly better than chance. If there are enough such similar persons, the scenes of the test person might be classified correctly to the some reasonable degree, despite the large number of dissimilar people. However, the experimental results listed in row *G-AAM-C* of Tab. 6.8 show that this hope is not fulfilled for the people in the object-teaching database. The average classification rate of 54.5% is only insignificantly better than chance.

Fusion of Classification Scores

As an improvement of the classification scheme used above, we took into account the "confidence" of the single classifiers. Instead of fusing the classification results, we fused the classification scores by taking the sum of them (please refer to Sec. 6.1.3). This increased the average classification accuracy to 62.2%, which, however, is still not good enough for practical applicability. Only for five subjects more than two out of three scenes were classified correctly on average.

Fusion of Reference Subsequences

The third and last fusion scheme we tested is a fusion of all reference subsequences of the single classifiers into a single, large set of reference subsequences. The resulting set is used in a single classification to assign a class label to the test scene.[14] Unexpectedly, this fusion

[14] As classification parameters for the resulting classifier, the median values of the individual classifiers' parameters (resp. geometric mean in case of the classification bias) are used, likewise to the experiments reported

scheme yielded the worst results of all: 51.6% accuracy on average, 4.6% for *success* and 97.7% for *failure*. Thus, the classifier constructed in this way completely failed to discriminate the two classes.

Conclusion

The negative expectation expressed at the beginning of the section has been confirmed: a person-independent classification yielded poor results, thus the classifiers trained with the dynamic recognition approach do not generalize to other people of the object-teaching study. The best average classification accuracy was 62.2% which is too poor for practical usage. A closer inspection of the classification results of the single models unveiled that it is actually typical that some models are strongly biased towards *failure* and others towards *success*. There are also a few cases were models perform approximately at random, but the biased ones are in majority. Thus, it is not unrealistic that these models will cancel out each other on average, so this part of the motivational reasoning above is (roughly) satisfied. However, what is completely missing is a sufficient number of "similar" people whose models perform well enough to dominate the poor classifications of the first group of models. Even in case one takes only the best-performing model to classify a test person, the average classification accuracy increases only to 67.5%. Thus, as even the most similar persons are not similar enough—in the sense discussed above—there is no real chance for the investigated fusion schemes to yield satisfactory results.

One obvious reason for this negative result is the comparatively small number of 11 people in object-teaching database, which appears to be not enough to cover a sufficient range of different behaviors regarding FCS display that would be necessary to generalize to new persons. However, in case the number of training people could be substantially increased, these fusion schemes face another problem: as they classify the test person with the model (or the reference subsequences) of each training person, the computational cost is linear in the number of persons, which is problematic for a real-time classification. Thus, the presented fusion schemes remain tentative in nature, but might serve as a first baseline for further investigations of the generalization to new persons in future research nevertheless.

6.3.4. Constrained Local Models for Feature Extraction

In the experiments reported in this section, the AAM that performed the feature extraction so far is replaced by a constraint local model (CLM) (please refer to Sec. 4.5.2). Likewise to the AAM, the model parameters of the CLM are used as feature vectors. We used the CLM implementation of the *FaceTracker* [154] software developed by Saragih, which implements a modified version of the method presented by Saragih *et al.* [440]. This C++ library is wrapped into a plugin for our *iceWing* framework [337] which we also used for the experiments so far. Thus, the AAM features can easily and transparently be replaced by the CLM features. In the experiments, we used the pretrained generic CLM that is provided with the software.

Table 6.9 shows the results of these experiments. In a person-specific classification, the average classification accuracy was 68.3% which is notably lower than the accuracy of the individual AAMs and also slightly lower, but nevertheless comparable, to the classification accuracies

in Sec. 6.3.1. An exception is the overall sum of reference subsequence number, which is the sum of the individual numbers.

features	01	02	03	04	05	06	07	08	09	10	11	mean	SD
– all	61	66	62	83	66	71	54	69	66	63	90	68.3	10.2
CLM – success	60	71	72	89	56	67	44	73	50	58	88	66.2	14.4
– failure	61	58	44	76	75	77	61	64	74	67	91	68.0	12.5
– all	88	82	80	92	88	89	72	86	85	92	95	86.3	6.5
T-CLM – success	84	84	88	100	90	88	69	92	71	88	89	85.7	8.9
– failure	92	78	66	83	84	90	75	79	92	94	99	84.7	9.7
– all	39	52	53	69	50	39	45	54	43	42	37	47.6	9.4
CLM-C – success	33	76	79	84	75	60	76	81	92	75	79	73.8	15.4
– failure	44	17	11	53	25	15	19	20	17	8	9	21.8	14.4
– all	45	52	75	78	78	39	61	63	40	33	63	57.0	16.1
CLM-S – success	33	24	100	84	94	60	72	78	100	58	71	70.4	25.3
– failure	65	92	33	71	63	15	52	44	9	8	57	45.4	26.7
– all	70	48	74	81	75	46	70	58	63	46	73	64.0	12.5
CLM-R – success	47	18	62	68	81	40	56	28	42	17	42	45.5	20.4
– failure	89	92	94	94	69	54	81	96	74	75	94	82.9	13.7

Table 6.9.: Classification accuracies using generic CLMs. Row *CLM* shows the results for a person-specific classification, while row *T-CLM* lists the respective mean leave-one-out training errors. Row *CLM-C* shows the results for a person-independent classification using the fusion of classification results scheme, row *CLM-S* the results for a fusion of classification scores, and row *CLM-R* the results for a fusion of reference subsequences. For each person, the classification accuracy for all scenes, only *success*, and only *failure* scenes is shown, as well as the mean accuracy and standard deviation over all persons. Please refer to Sec. 6.3.4.

of the generic AAMs. The correlation between the leave-one-out training error and the test performance is significant (Spearman correlation, $\rho \approx 0.65$, $p < 0.03$).

In a person-independent classification using the fusion of classification results and fusion of classification scores schemes, the CLM also performed slightly worse than the AAMs. Surprisingly, the best results were achieved by a fusion of reference subsequences, which failed completely in the AAM case. However, despite yielding the best generalization result overall, the classification accuracy of 64.0% on average is still not good enough for practical usage.

6.4. Conclusion

In this chapter, we presented a dynamic recognition approach for the classification of facial communicative signals and evaluated it on the videos of the object-teaching database (please refer to Sec. 3.2). In a person-specific classification with individual AAMs, the approach achieved a good average classification accuracy of 81.9%, which basically reached the average human performance of 82.0% (please see Sec. 3.5.2. Likewise to the human classification, the variance over different persons was very high. The achieved classification performance is notably better than the best one of the static baseline approach evaluated before (76.0%, please see Sec. 5.4.2).

It turned out that stable and well-suited classifier parameters exist for each person. However, these parameters are specific to the respective person and do not generalize well to other people. An inspection of the selected reference subsequences showed that all three kinds of

FCSs we focused our investigations on—head gestures, eye gaze, and facial expressions—were selected as reference subsequences. The reference subsequences for the individual persons reflect typical behavior for them as evaluated in Sec. 3.4. There are in part large differences between the best reference subsequences of different persons.

Unfortunately, the usage of generic AAMs in a person-specific classification reduced the classification accuracies considerably. Compared to the results for individual AAMs, the differences between generic models with and without the target person were marginal (72.7% vs. 70.3% on average). The results for a generalization to new people using a person-independent classification were all negative. The best average classification accuracy of 64.0% was achieved by an generic CLM. However, this is to poor for practical usage.

Taking these results together, the only well-suited application scenario for the dynamic FCS recognition approach appears to be a person-specific classification with individual AAMs. Please refer to Sec. 7.1.1 for an outline of how this might be done in practice. However, it is also not clear to which degree the humans performed a person-dependent or -independent classification. As the subjects did not know the people shown in the videos before the experiments, they relied on their "world knowledge" in the beginning. Nevertheless, our impression was that the subjects rapidly became more familiar with some individual characteristics of the shown people as the experiment proceeded. (Some subjects also verbally hinted this during the experiment.) The subjects could watch each scene as often as they wanted before moving on to the next scene. Thus, we conjecture that the subjects started doing a person-independent classification in the beginning and performed an increasingly person-dependent classification as the experiment proceeded, at least a certain adaptation to the shown people took place. However, this could not be measured in a formal or quantitative way.

7. Conclusion

What we know is a drop, what we don't know is an ocean.
— Isaac Newton

In this dissertation, we investigated facial communicative signals (FCSs) in a valence recognition scenario in task-oriented human-robot interaction. The first chapter provided an introduction and motivated this work. In the second chapter, we discussed some traits of the human face perception in general and reviewed the psychological research on the human display and perception of FCSs, in particular head gestures, eye gaze, and facial expressions. Our definition of FCSs is pragmatic and focuses on the attribution of meaning, not on the visual appearance of a facial display. We drew several conclusions that motivated the work presented in the subsequent chapters. One of these conclusions is that it is necessary to investigate FCSs in specific interaction scenarios and thus within certain contexts which also characterize the scope of validity of the found results, as a FCS investigation "in general" in single interaction scenarios is not feasible due to the complex nature and high context-dependence of FCS.

In the third chapter, we introduced the object-teaching scenario as a concrete example of a task-oriented human-robot interaction situation. We recorded a video database showing 11 subjects interacting with a robot by teaching the names of several objects to it. We formulated a valence-based approach for FCS recognition where facial displays are interpreted in terms of *success* and *failure*. The ground truth for these facial displays is defined by the interaction situation, namely the correct or incorrect verbal answer given by the robot, in contrast to a definition based on the visual appearance which is used in most investigations of FCSs. We evaluated the facial displays of the 11 subjects during their reaction to the robot, which unveiled in part large differences in the communicative behaviors of these people.

Furthermore, we investigated the human recognition performance in this scenario. The FCSs shown in the interaction scenes were judged by 44 new subjects, who decided whether the robot's answer was correct or not, based on the observed facial displays. This evaluation was conducted under different visual and temporal context conditions. It turned out that humans can correctly interpret the FCSs to a reasonable degree even if only the face is shown without visual context, given a sufficient length of the observed video sequence; the visual context shown in the videos is not required for this interpretation. Thus, automatic recognition approaches can reasonably focus on the face only. However, the average classification accuracy of the humans was only 82.0%, which is comparatively low for a classification problem with two classes and demonstrates the difficulty of this task. The variance over both the shown videos and the observing subjects was very high.

The fourth chapter presented an overview of state of the art approaches for an automatic recognition of FCSs. We discussed methods for face detection and the recognition of head gestures, eye gaze, and facial expressions. The *Encara* face detection approach (Sec. 4.1.6), active appearance models (AAMs, Sec. 4.5.1), constrained local models (CLMs, Sec. 4.5.2),

and Gabor energy filters (GEFs, Sec. 4.5.3) were considered in some more detail because the were utilized in our investigations of automatic FCS recognition in chapters five and six.

In the fifth chapter, we considered a simple static recognition approach using support vector machines (SVMs). This approach did not make use of any temporal dynamics in the videos, but used a majority voting over the classification results of the single frames of a video. This simple majority voting scheme turned out to be not beneficial for this classification task, as a classification where each video was represented by the mean feature vector of its frames only yielded comparable, even slightly better results. An investigation of this finding led to the hypothesis that only comparatively short subsequences of the scenes are actually discriminative in terms of *success* and *failure*, despite the pre-segmentation of the videos. The best average classification accuracy was 76.0%, achieved by a feature extraction with AAMs. The results of the static recognition approach served as a baseline for the dynamic approach presented in the sixth chapter.

The sixth chapter described our dynamic recognition approach in detail. Motivated by the results of the static baseline approach, visual inspection of the interaction videos, and the general finding that the temporal dynamics appear to be important for FCS interpretation, we developed a classification approach based on discriminative reference subsequences and dynamic time warping (DTW). We explained the approach in detail and considered also a prototype implementation in Matlab. Furthermore, we discussed several related approaches where we focused especially on the applied pruning techniques and the difficulties in transferring them to our approach.

In the evaluation on the videos of the object-teaching database, the dynamic recognition approach outperformed the static one and yielded an average classification accuracy of 81.9% in a person-specific classification using AAMs and thus reached the average human performance. Likewise to the human classification, the variance over different persons was very high. Well-suited classifier parameters turned out to be person-specific. In tentative experiments regarding a person-independent classification, the achieved results were rather low. The best average classification accuracy was only 64.0% using CLMs. A main reason for this poor performance is the high variability of the communicative behaviors of the different persons. Paired with the comparatively low number of 11 subjects in the database, the classifier training could not cover a large enough amount of variance to generalize to new people. Thus, the effective application domain for the dynamic recognition approach in the current state of affairs is the person-specific classification.

7.1. Possible Application Scenarios

This section briefly sketches two possible application scenarios for the dynamic recognition approach for FCS classification in terms of valence. The first one is a realistic household robot scenario where a social robot assists a family in the daily life. The second one is more tentative and concerns the question how human teachers can effectively support the training of an artificial intelligent agent using reinforcement learning.

7.1.1. A Social Robot at Home

Imagine a social robot that supports a family at their home in the activities of daily life, for instance by cleaning the rooms, preparing meals, or being a companion for games. Regarding

FCSs, the robot is especially supposed to recognize the facial displays of the family members in typical interactions that come along with his tasks. Therefore, individual AAMs for the family members are built by the robot manufacturer based on several minutes of recorded video material of each person. At the beginning of a new interaction, the robot performs a face identification to load the personalized model for its current interaction partner.[1] Based on the instructions given by the human (e.g. "I want to show you some new objects." or "Help me with the cooking."), the robot determines to which scenario the just started interaction belongs and selects the respective FCS classifier trained for this person in this scenario.

The training of these classifiers might be performed in the following way. It is assumed that in each scenario a criterion for *success* or *failure* of the performed actions exists which is normally ascertainable by the robot. Usually, like in case of the object-teaching scenario, this can be the verbal reactions of the human to the answers or actions of the robot, confirming or discarding its behavior. Whenever the speech recognition system of the robot recognized such a verbal reaction with high confidence, it is used as ground truth label for the respective facial displays occurring in this interaction. The training data accumulated in this way is used to retrain the classifier in the background or at certain intervals, for instance over night.

The trained FCS classifiers for valence recognition are in turn utilized when the confidence of the speech recognition is too low, for instance when music playing in the background or other noises disturb the robot's speech perception, or if the human does not give a verbal reaction. In this case, the robot tries to get feedback about its actions from the perceived FCSs.

The addition of new persons to the database of the robot might be supported be a semi-automatic bootstrapping approach. This assumes that a large generic AAM comprising images from many different persons with high variance preservation is available. This AAM is applied to video recordings of the new person offline, in order to produce as many good matchings as possible. As the fitting is done offline, there is no need for real-time performance, thus a very exhaustive search can be performed to enhance the fitting quality. While still a large number of video frames might have a poor fitting, a significant number of frames is expected to feature a fitting of sufficient quality nevertheless. Using a threshold on the reconstruction error, the latter images—together with the fitted feature points—are used to build an individual AAM for the new person, either fully automatically or after an inspection and verification of a human expert, possibly by adding further, hand-annotated images of the person showing facial displays that were poorly covered by this automatic bootstrapping method. Even in the latter case, the human effort would be significantly reduced by this approach.[2]

In the investigations presented in this thesis, presegmented videos were used. Of course, the social robot in the scenario considered here has to perform this segmentation online and automatically. However, this is not a critical issue, a simple sliding window approach with a fixed, predefined maximal window length can be used. A new frame can be added to the input sequence efficiently: according to Bellman's principle of optimality [29], the dynamic programming tables used to compute the dynamic time warping during the classification do not need to be recalculated, but just one more column is added to each one (please refer to Alg. A.2 and A.7). The oldest frame of the input video sequence is dropped when a new one arrives, the minimal distances of reference subsequences whose matched subsequence started at the dropped frame are also rejected. However, this procedure requires an additional

[1] In previous work [304, 207], we successfully used AAMs also for face identification. However, many other methods are available for this task (e.g. [68, 537, 411]).

[2] We applied such a bootstrapping method for the AAM training in previous work [304, 207].

space linear in the maximal length of the reference subsequences, because several dynamic programming tables need to be kept in memory in parallel.

7.1.2. Supporting the Training of Artificial Agents using Reinforcement Learning

Knox and Stone [289] presented a method called *TAMER* to train a virtual agent by positive and negative feedback of a human using reinforcement learning [263]. They demonstrated a very good performance in experiments where the virtual agent learned to play the game "Tetris" from the feedback of humans who observed the way of playing of the agent, judging its decisions either positively or negatively. The virtual agent could play Tetris reasonably well after three games of training, which is significantly faster than automatic training approaches.

An idea for an attempt to enhance this human teaching is to gain the human feedback from the interpretation of the facial displays in terms of valence, so that the human teacher does not need to press buttons or to do something similar. For the Tetris scenario, this is probably not needed as the training is very efficient already. Nevertheless, their might be other scenarios where this would lead to a real improvement, for instance if pressing buttons, etc. is not suitable due to the particularities of the used setting. However, the humans would probably have to pose the FCSs to some degree to improve the automatic recognition, because probably a very high accuracy is required in this case for the reinforcement learning to be efficient.

7.2. Outlook

In the final section of this dissertation, we outline some promising directions for future research. We consider concrete suggestions how the dynamic recognition approach might be improved first, before we state a few more general research directions that are worthy to be investigated.

7.2.1. Improving the Dynamic Recognition Approach

There are several possibilities to improve and extend the dynamic approach for FCS recognition that can be investigated in future research. The selection of reference subsequences might be improved by incorporating additional criteria, besides featuring a high discriminativity-value. For instance, it can be required that two reference subsequences of the same class may not be too similar to each other to avoid a "waste" of prototypes by covering the same area of the class space twice. As discussed in Sec. 6.1.2, a subsequence with high discriminativity-value might not necessarily yield a good classification accuracy: it could be the case that this subsequence is matched rarely, thus it classifies a small number of test sequences very well, but does not contribute to the majority of classifications. To cope with this issue, a certain number of reference subsequences are preferably selected to improve the coverage of the class space. However, this might be more directly addressed by estimating the number of expected matches during classification, based on the statistics of the training data, and using this as additional selection criterion.

In Sec. 6.2.1, we discussed a pruning strategy based on the *LB_Keogh* lower bounding technique [282, 280] which might lead to some speedup in the classification procedure of our approach. This could be empirically evaluated in future work. Furthermore, the allowed dynamic time

warping path might be restricted to certain areas of the matrix in general to reduce the computational cost. The studies conducted by Xi *et al.* [526] showed very promising results for such an approach.

In the course of developing the dynamic recognition approach, we also experimented with other machine learning techniques, including hidden Markov models, Gaussian mixture models, AdaBoost, particle filters (condensation trajectory recognition [222, 36]), and non-negative sparse coding [228]. However, in our tests, none of these methods did outperform the static baseline approach. One aspect of the problem appears to be that the object-teaching database does not contain enough data per person to build appropriate statistical distribution models. Thus, prototype-based methods, like the dynamic recognition approach, seem to be better suited also for that reason, according to present knowledge.

Sticking with a prototype-based classification, an interesting idea for future research might be to replace the nearest-neighbor-based classification by a SVM classification, as SVMs usually yield better results than NN classifiers. The main problem here is that the standard SVM approach cannot be used for time series data of varying size with a DTW distance. Nevertheless, interesting approaches which try to incorporate DTW or similar elastic distance measures into SVMs have been proposed. Shimodaira *et al.* [457] argued that DTW can be used as a kernel function for SVMs. Unfortunately, several years later, Lei and Sun [316] proved theoretically that the DTW distance is not positive definite symmetric and thus not a valid kernel. Despite this result, Chaovalitwongse and Pardalos [72] achieved good empirically results using it as a kernel. Zhang *et al.* [552] used a Gaussian kernel with two elastic distance measures that are similiar to DTW, but are proper metrics. They could not prove that the resulting kernels are actually positive definite symmetric, but in all their experiments, this was the case in practice. Their method yielded good results on many datasets. Chen and Ng [75] introduced the *edit distance with real penalty* which also has favorable properties. Investigations along these lines appear to be promising for a further development of the dynamic recognition approach in future work.

7.2.2. Interesting Questions for Future Research

Besides improvements of the core elements of the dynamic recognition approach as discussed in the previous section, there are many other research questions that are worthy to be investigated in future work. The detailed investigation of further interaction scenarios and their comparison to the object-teaching scenario would be an important step towards a more comprehensive understanding of the nature and specificity or generality of the involved FCSs. Also the consideration of an additional *neutral* class might be beneficial for some interaction scenarios. Similarly, the rejection of interaction scenes where the classifier is not confident enough could be investigated, as no classification is better than a wrong classification in many cases.[3]

Section 6.3.1 presented a very remarkable result: the *AAM-96* features performed considerably different than the *AAM-99* features regarding the distribution of classification accuracies over the individual persons, but in fact significantly similar to the static approach. This surprising

[3] An obvious idea for such an rejection is to use a threshold on the differences of the classification scores and to reject a scene in case the absolute value of this difference is too small. However, preliminary evaluations suggest that this simple method is not well-suited, as often the classification accuracy on the non-rejected data does not improve, despite rejecting a possibly large number of scenes.

finding should be investigated in greater detail, in order to understand how this result relates to specific characteristics of the respective persons. In the best case, this might finally lead to an automatic selection of person-specific features and/or classification methods which are well-suited for this person, which would yield a notable enhancement of the classification performance.

Though the results regarding a person-independent classification presented in this thesis are negative, a reasonable generalization to new persons remains an important—and very challenging—target for future research. Further investigations for a substantially larger number of people are necessary in order to better understand what exactly are the problems and how they can be adequately addressed. According to the observation stated in Sec. 6.4, the human subjects who classified the object-teaching videos seemed to rapidly adapt to the shown person. Hence such an online adaptation to previously unknown people might be a key issue for the desired generalization.

A. Appendix

A.1. Instructions for the Subjects of the Object-Teaching Study

The subjects received the following written instructions and could also ask questions before the experiment started (english translation below):

Vielen Dank für Ihre Teilnahme am Experiment. Im Folgenden erhalten Sie genauere Informationen dazu.

Auf diesem Bild sehen sie BIRON. Bei der Entwicklung des Roboters streben wir einen Assistenten im Haushalt an, das heißt, er soll in der Lage sein, sich in einer Wohnung zu orientieren und verschiedene Aufgaben des täglichen Lebens zu erledigen. Um sich in seiner Umgebung zurechtzufinden, muss der Roboter lernen können. Ziel des Experiments ist es deshalb, BIRON verschiedene Objekte zu zeigen. Sie werden während der Interaktion mit dem Roboter eine Auswahl neben sich auf einem Tisch finden. Bitte beachten Sie, dass BIRON die Objekte vor der Interaktion mit Ihnen nicht kennt und sie erst lernen muss. Bitte zeigen Sie BIRON nacheinander einige Objekte und bringen Sie ihm deren Namen bei. Bitte überprüfen Sie auch, ob er sie tatsächlich gelernt hat.

Wichtig beim Gespräch mit BIRON:

1. Bitte bedenken Sie, dass der Roboter manchmal etwas mehr Zeit für die Verarbeitung braucht.
2. Bitte versuchen Sie während des Experiments alle Probleme mit dem Roboter und nicht mit den anderen anwesenden Personen zu lösen.
3. Damit BIRON und die Kameras Sie nicht “aus den Augen verlieren”, ist es wichtig, dass Sie möglichst die ganze Zeit unmittelbar vor dem Tisch stehen.

Der Ablauf des Gesprächs sollte darüber hinaus wie mit einem Menschen auch erfolgen (begrüßen sie BIRON, unterhalten Sie sich mit ihm, verabschieden Sie sich von ihm). Bitte führen Sie das Experiment in zwei Phasen durch, die Objekte werden zwischendurch einmal ausgetauscht. Sie haben für jede Phase ungefähr 10 Minuten Zeit. Nach dem Experiment bitten wie Sie, noch einen kurzen Fragebogen auszufüllen.

Sie werden während des Experiments gefilmt. Sie können das Experiment jederzeit ohne Angabe von Gründen abbrechen oder Teile bzw. Fragen auslassen. Nach dem Experiment erhalten Sie sechs Euro Aufwandsentschädigung.

Viel Spaß!

Englisch translation:

Thank you very much for your participation in this experiment. Below you find precise information about it.

On this picture you see BIRON. In developing this robot we aim at a household assistant, this means it shall be able to orientate itself in an apartment and to handle various issues of daily life. To find its way in its environment, the robot must be able to learn. Therefore, the goal of the experiment is to show several objects to BIRON. During the interaction with the robot, you will find a selection on the table besides you. Please note that BIRON does not know the objects before the interaction with you and needs to learn them first. Please successively show some objects to BIRON and teach it their names. Please also check whether it actually learned them.

Important issues for the conversation with BIRON:

1. Please note that the robot sometimes needs a little more time for processing.
2. During the experiment, please try to resolve all issues with the robot and not with other persons present.
3. To ensure that BIRON and the cameras do not “lose sight of you”, it is important that you stay close to the table for the whole time as far as possible.

Additionally, the course of conversation should follow a human conversation (greet BIRON, talk to him, and say good bye). Please perform the experiment in two phases, the objects will be exchanged in between. You have got about 10 minutes time for each phase. After the experiment, we ask you to fill out a short questionnaire.

You will be recorded on video during the experiment. You may cancel the experiment any time without giving reasons or omit parts resp. questions. After the experiment, you will receive six euro compensation.

Have fun!

A.2. Utterances of Biron in the Object-Teaching Study

Greeting:

- Hallo. *(Hello.)*
- Mein Name ist Biron. *(My name is Biron.)*

Human interaction partner shows object and tells its name (object learning):

- Das ist also ein[e] `<object_name>`. *(So, this is a `<object_name>`.)*

Human interaction partner asks Biron for the object name (object recognition):

- Das ist ein[e] `<object_name>`. *(This is a `<object_name>`.)*

Utterances Biron used now and then, particularly when the human interaction partner taught or inquired an object name:

- Wie bitte? *(Pardon?)*
- Ich konnte Dich nur teilweise verstehen. Kannst du das bitte noch einmal wiederholen? *(I could merely understand you partially. Can you repeat this, please?)*
- Hast Du mir das Objekt schon einmal gezeigt? *(Did you show me the object before?)*
- Das ist interessant. *(This is interesting.)*
- Es gefällt mir gut. *(I like it.)*
- Ich kenne das Wort nicht. *(I don't know the word.)*
- Ich kenne das Objekt nicht. *(I don't know the object.)*
- Ich weiß es nicht. *(I don't know.)*
- Ich sehe kein Objekt. *(I don't see any object.)*
- Ich habe das Objekt gesehen. *(I have seen the object.)*
- Nein. *(No.)*
- Ja. *(Yes.)*
- Vielleicht versuchen wir etwas anderes? *(Maybe we try something else?)*

Farewell:

- Auf Wiedersehen. *(Good bye.)*
- Es hat mir großen Spaß gemacht. *(This was a lot of fun.)*

Object names Biron used:

- Ball, Tasse, Lineal, Tuch, CD, Kugelschreiber, Schokoriegel, Flasche, Schere, Handy, Kugel, Stift, Buch, Riegel, Kuli, Tasche, Mars, DVD, Süßigkeit, Schallplatte, Mars-Schokolade, Becher, Schokolade, Snickers, Bier, Bierflasche *(ball, mug, ruler, cloth, CD, ballpen, chocolate bar, bottle, scissors, mobile phone, sphere, pen, bar, ballpoint pen, bag, Mars, DVD, candy, disc, Mars chocolate, cup, chocolate, Snickers chocolate, beer, beer bottle)*

A.3. Questionnaire for the Subjects of the Object-Teaching Study

1. Kannten Sie BIRON bereits vor dem Experiment?
   (Did you already know BIRON before the experiment?)
2. Haben Sie schon einmal an einem Experiment mit BIRON teilgenommen?
   (Did you ever participate in an experiment with BIRON before?)
3. Kennen Sie andere Roboter oder haben Sie schon einmal an Experimenten mit anderen Robotern teilgenommen oder mit ihnen gearbeitet?
   (Do you know other robots or have you ever participated in experiments with other robots or did you work with them?)
4. Wie gut hat BIRON die Objekte gelernt?
   (How well did BIRON learn the objects?)
5. Wie haben Sie die Interaktion mit BIRON insgesamt empfunden?
   (How did you perceive the interaction with BIRON all in all?)
6. Wie autonom hat BIRON ihrem Eindruck nach agiert?
   How autonomously did BIRON act, according to your impression?)
7. Haben Sie weitere Anmerkungen zum Experiment?
   (Do you have additional remarks to the experiment?)

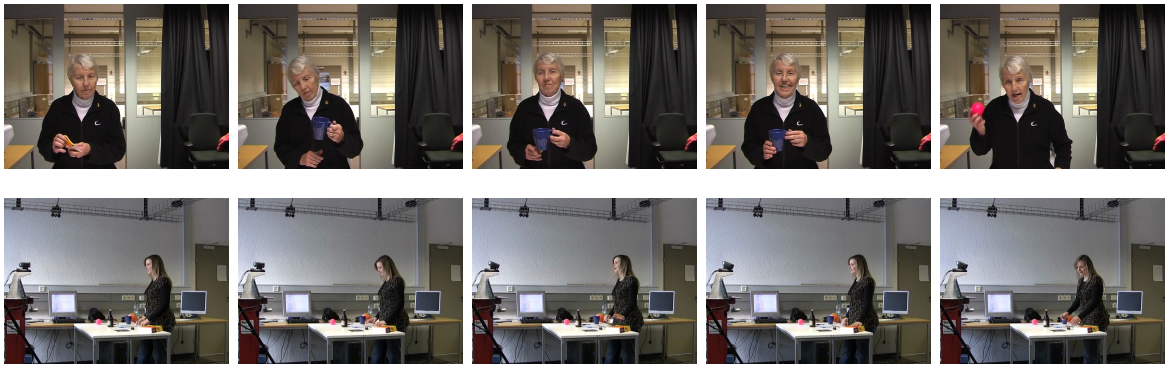
A.4. Instructions for the Subjects of the Valence Recogniton Study

The subjects received the following instructions and could also ask questions before the experiment started (english translation below):

Die bei dieser Studie gezeigten Videos zeigen Versuchspersonen bei der Interaktion mit dem Roboter Biron. Die Versuchspersonen zeigen Biron verschiedene Objekte und versuchen ihm deren Namen beizubringen. Die Interaktion verläuft üblicherweise in vier Phasen:

1. Die Versuchsperson zeigt dem Roboter das Objekt und sagt den Namen (oder fragt danach)
2. Die Versuchsperson wartet auf die Antwort des Roboters
3. Der Roboter antwortet und sagt dabei möglicherweise den richtigen oder einen falschen Objektnamen
4. Die Versuchsperson reagiert auf die Antwort des Roboters

Two example video sequences are shown to make the subjects familiar with the kind of videos they are about to judge. The first one shows a 63 seconds long (continuous) part of an interaction from the perspective of the stationary face camera, covering several interaction scenes (in the sense of the human valence recognition study reported in Sec. 3.5). The second video sequence shows a 89 seconds long (continuous) part of another interaction from the perspective of the stationary scene camera:



In der Studie verwenden wir nur Videos aus der ersten Perspektive. Sie zeigen Interaktionen, die entweder positiv verlaufen sind – d.h., der Roboter hat in der dritten Phase den richtigen Objektnamen gesagt – oder Interaktionen, die negativ verlaufen sind – d.h., der Roboter hat in der dritten Phase einen falschen Objektnamen gesagt. Die Videos beginnen jeweils am Ende der dritten Phase – in dem Moment, wo der Roboter den Namen des Objektes sagt – und laufen bis zum Ende der vierten Phase (das Ende hat ein Annotierer festgelegt). Die Videos enthalten keinen Ton und verschieden viel "Kontext".

Ihre Aufgabe ist es jetzt, für jedes Video einzuschätzen, ob die gezeigte Interaktion Ihrer Meinung nach positiv oder negativ war.

[Some technical details about how to fill out the questionnaire follow.]

Viel Spaß und vielen Dank!

Englisch translation:

The videos shown in this study display subjects during an interaction with the robot Biron. The subjects show several objects to Biron and try to teach it the objects' names. An interaction usually proceeds in four phases:

1. The subject shows an object to the robot and tells the name (or asks for it)
2. The subject waits for the answer of the robot
3. The robot answers and possibly says the correct or a wrong object name
4. The subject reacts to the answer of the robot

In the study, we only use videos from the first perspective. They show interactions that either proceeded positively—i.e., the robot said the correct object name in the third phase—or interactions, that proceeded negatively—i.e., the robot said a wrong object name in the third phase. Each video starts at the end of the third phase—at the momemnt when the robot says the object name—and plays until the end of the fourth phase (this end was determined by an annotating person). The videos do not contain sound and show different amounts of "context".

Now, your task is to judge for each video whether the displayed interaction was positive or negative, according to your estimation.

Have fun and thank you very much!

A.5. Questionnaire for the Subjects of the Valence Recogniton Study

First, the subjects performed the study where they were asked to make a forced choice for every video. Afterwards, they were asked to answer to following questions:

1. Welche Merkmale in den Videos haben Sie für Ihre Einschätzung vorrangig benutzt bzw. worauf haben Sie besonders geachtet?
   (Which features in the videos did you primarily use for your evaluation resp. to what did you pay attention in particular?)
2. Haben Sie noch irgendwelche Anmerkungen zu diesem Experiment?
   (Do you have any additional remarks regarding this experiment?)

A.6. Implementation of the Dynamic Recognition Approach

The following algorithms Alg. A.1 to Alg. A.12 outline our prototype implementation of the dynamic recognition approach. In these algorithms, the following variables are frequently used and assumed to be accessible, although only the main parameters are listed in the input parameter lists to enhance clarity:

- N ... number of training videos
- N_{success} ... number of training videos of class *success*
- N_{failure} ... number of training videos of class *failure*
- L_i ... length of the i-th training video in frames
- A ... multi-dimensional matrix containing the AAM parameter vectors for all frames of all training videos:
  - A_i ... AAM parameter vector sequence of the i-th video
  - A_i^j ... AAM parameter vector of the j-th frame in the i-th video
- c_i ... class of the i-th training video (*success* or *failure*)
- a ... AAM parameter vectors of all frames of a test video:
  - a_j ... AAM parameter vector of the j-th frame in the test video
- L_a ... length of a test video in frames

Furthermore, the following variables represent the various parameters of the approach (please see Tab. 6.1):

Algorithm A.1 Perform training, i.e. select reference subsequences from given training data

function $(R_{\text{success}}, R_{\text{failure}}) \leftarrow$ perform_training (A, c, P)

// ***Input:***
// A ... *AAM parameter vectors of all training videos*
// c ... *class label for each training video*
// P ... *parameter values to evaluate in the grid search*
// ***Output:***
// $(R_{success}, R_{failure})$ ... *selected reference subsequences*

// Precomputation of frequently required distances → Alg. A.4 and Alg. A.6
$F \leftarrow$ precompute_frame_distances (A)
$S \leftarrow$ precompute_subsequence_distances $(A, F, \min\{P.L_{\min}\}, \max\{P.L_{\max}\}, \max\{P.K\})$
// Find classifier parameters that perform best on training data → Alg. A.3
$p \leftarrow$ perform_parameter_grid_search (A, c, F, S, P)
// Compute the reference subsequences → Alg. A.8 and Alg. A.9
$n_c \leftarrow p.v + \max\{0,\ p.t - N_c \cdot p.v\} \quad \forall\, c \in \{success, failure\}$
for $m \leftarrow 1$ **to** N
  $R_m \leftarrow$ select_reference_subsequences $(S, c, m, \{i \mid 1 \le i \le N, i \ne m\}, p.l_{\min}, p.l_{\max}, p.k, n_{c_m})$
end for
$(R_{\text{success}}, R_{\text{failure}}) \leftarrow$ fuse_reference_subsequences $(\{R_m \mid 1 \le m \le N\}, c, p.v, p.t)$
return $(R_{\text{success}}, R_{\text{failure}})$

end function

- f ... subsequence length difference pruning factor (please refer to Eq. 6.2)
- $P.L$ ... set of pairs $(l_{\min}, l_{\max})$ of subsequence length parameters
- $P.L_{\min}$ ... alternative notation for a set containing the $l_{\min}$-parameters of $p.L$ only
- $P.L_{\max}$ ... alternative notation for a set containing the $l_{\max}$-parameters of $p.L$ only
- $P.K$ ... set of k parameters
- $P.T$ ... set of t parameters
- $P.V$ ... set of v parameters
- $P.U$ ... set of u parameters
- $P.W$ ... set of w parameters

Thus, P holds the parameter values that are used for parameter optimization in the grid search. The optimal parameters found during this search are denoted by $p.l_{\min}$, $p.l_{\max}$, $p.k$, $p.v$, $p.t$, $p.u$, $p.w$, and $p.b$, where $p.b$ is the classification bias (please see Sec. 6.1.4).

A.7. Feature Vector Dimensionality

The dimensionality of the various feature vectors that were used for the classification is listed in Tab. A.1.

Algorithm A.2 Perform classification, i.e. classify a test video

function $c \leftarrow$ perform_classification (R_{success}, R_{failure}, a, p)

*// **Input:***
// $R_{success}, R_{failure}$ ... selected reference subsequences → Alg. A.1
// a ... AAM parameter vector sequence of the test video
// p ... parameters of the classifier
*// **Output:***
*// c ... assigned class label (*success *or* failure*)*

$D^* \leftarrow$ compute_minimal_distances_to_reference_subsequences (R_{success}, R_{failure}, $\emptyset$, a) *// → Alg. A.7*
$(d_{\text{success}}, d_{\text{failure}}) \leftarrow$ compute_classification_scores (D^*, $p.t$, $p.u$, $p.w$) *// → Alg. A.10*

$$c \leftarrow \begin{cases} success & | \text{ if } d_{\text{success}} \geq p.b \cdot d_{\text{failure}} \\ failure & | \text{ if } d_{\text{success}} < p.b \cdot d_{\text{failure}} \end{cases}$$

return c

end function

feature extraction model	Sec.	01	02	03	04	05	06	07	08	09	10	11
AAM-95	5.4.1	27	33	37	33	27	38	29	35	23	37	26
AAM-99	5.4.1	67	73	78	74	64	75	69	73	59	75	58
GEF-4	5.4.1	640 for all persons										
GEF-8	5.4.1	2,560 for all persons										
GEF-12	5.4.1	5,760 for all persons										
RGB-8	5.4.1	192 for all persons										
RGB-16	5.4.1	768 for all persons										
RGB-25	5.4.1	1,875 for all persons										
gray-8	5.4.1	64 for all persons										
gray-16	5.4.1	256 for all persons										
gray-25	5.4.1	625 for all persons										
AAM-96	6.3.1	33	40	44	40	32	44	35	41	29	43	31
AAM-99	6.3.1	67	73	78	74	64	75	69	73	59	75	58
G-AAM	6.3.2	93 for all persons										
L-AAM	6.3.2	94	92	91	92	93	91	93	91	93	93	92
CLM	6.3.4	28 for all persons										

Table A.1.: The dimensions of the feature vectors for different feature extraction models (rows), in each case for the subjects of the object-teaching study (columns). Please refer to the indicated sections for details on the conducted classification experiments.

Algorithm A.3 Perform parameter grid search to find the best classifier parameters

function $p \leftarrow$ perform_parameter_grid_search (A, c, F, S, P)

*// **Input:***
// A ... AAM parameter vectors of all training videos
// c ... class labels of all training videos
// F ... precomputed frame distances (→ Alg. A.4)
// S ... precomputed subsequence distances (→ Alg. A.6)
// P ... parameter grid values to evaluate (→ Sec. 6.1.6)
*// **Output:***
// p ... best performing parameters (→ Tab. 6.1):
// $p.l_{min}$, $p.l_{max}$, $p.k$, $p.v$, $p.t$, $p.u$, $p.w$, $p.b$

// Initialization
$Z \leftarrow \emptyset$
$n_c \leftarrow \min\{P.v\} + \max\{\, 0,\ p.t - N_c \cdot \min\{P.v\} \,\} \quad \forall\, c \in \{success, failure\}$
// Grid search over reference subsequence selection parameters
for $(l_{\min}, l_{\max}, k) \in P.L \times P.K$
  // Trial classify each training video (leave-one-out optimization)
  for $m_1 \leftarrow 1$ **to** N
    // Compute reference subsequences for classification of the m_1-th video
    for $m_2 \leftarrow 1$ **to** N **with** $m_2 \neq m_1$
      $M \leftarrow \{\, i \mid 1 \leq i \leq N, i \neq m_1, i \neq m_2 \,\}$
      $R_{m_2} \leftarrow$ select_reference_subsequences (S, c, m_2, M, $l_{\min}$, $l_{\max}$, k, $n_{c_{m_2}}$) *// → Alg. A.8*
    end for
    for $v \in P.V$
      // → Alg. A.9 and Alg. A.7
      $(R_{\text{success}}, R_{\text{failure}}) \leftarrow$ fuse_reference_subsequences ($\{R_m \mid 1 \leq m \leq N, 1 \neq m_1\}$,$c$,$v$,$\max\{P.T\}$)
      $D^* \leftarrow$ compute_minimal_distances_to_reference_subsequences (R_{success}, R_{failure}, F, m_1)
      // Grid search over classification parameters
      for $(t, u, w) \in P.T \times P.U \times P.W$
        $Y^{m_1}_{v,t,u,w} \leftarrow$ compute_classification_scores (D^*, t, u, w) *// → Alg. A.10*
      end for
    end for
  end for
  // Evaluate classification scores
  for $(v, t, u, w) \in P.V \times P.T \times P.U \times P.W$
    $Y \leftarrow \{\, (Y^m_{v,t,u,w}, c_m) \mid 1 \leq m \leq N \,\}$
    $Z_{v,t,u,w,1.0} \leftarrow$ compute_classification_rate_and_margin (Y, 1.0) *// → Alg. A.12*
    $b \leftarrow$ compute_classification_bias (Y) *// → Alg. A.11*
    $Z_{v,t,u,w,b} \leftarrow$ compute_classification_rate_and_margin (Y, b) *// → Alg. A.12*
  end for
  $(r, \psi) \leftarrow (r', \psi') \in Z_{v,t,u,w,b}$ with r' is maximal and ψ' is maximal for all maximal r' values
  $(v, t, u, w, b) \leftarrow$ index of (r, ψ)
  $Z \leftarrow Z \cup \{\, ((l_{\min}, l_{\max}, k, v, t, u, w, b), (r, \psi)) \,\}$
end for
// Select best parameters overall
$(r, \psi) \leftarrow (r', \psi')$ from all $((l_{\min}, l_{\max}, k, v, t, u, w, b), (r', \psi')) \in Z$ with r' is maximal and ψ' is maximal for all maximal r' values
$p \leftarrow$ index of (r, ψ), reformatted to structure with fields $p.l_{\min}$, $p.l_{\max}$, $p.k$, $p.v$, $p.t$, $p.u$, $p.w$, $p.b$
return p

end function

Algorithm A.4 Precomputation of the frame distances for each pair of training videos

function $D \leftarrow$ precompute_frame_distances (A)

```
// Input:
//     A ... AAM parameter vectors of all training videos
// Output:
//     D ... matrix of frame distances:
//         D_{i,j}^{k,l} is the Euclidean distance of k-th frame in i-th video to l-th frame in j-th video

for i ← 1 to N − 1
    for j ← i + 1 to N
        for k ← 1 to L_i
            for l ← 1 to L_j
                D_{i,j}^{k,l} ← |A_i^k − A_j^l|
                D_{j,i}^{l,k} ← D_{i,j}^{k,l}
            end for
        end for
    end for
end for
return D
```

end function

Algorithm A.5 Compute the DTW distance matrix for two subsequences

function $D \leftarrow$ compute_DTW_distance_matrix (F)

```
// Input:
//     F ... submatrix of the frame distance matrix: → Alg. A.4
//         F_{i,j} ... Euclidean distance of the i-th frame in the first subsequence to j-th frame in the
//         second subsequence
// Output:
//     D ... dynamic time warping (DTW) distance matrix:
//         D_{i,j} ... DTW distance of the subsequence formed by the first i frames of the first input
//         subsequence to the subsequence formed by the first j frames of the second input
//         subsequence

D_{1,1} ← F_{1,1}
for i ← 1 to maximal index i in F
    D_{i,1} ← D_{i−1,1} + F_{i,1}
end for
for j ← 1 to maximal index j in F
    D_{1,j} ← D_{1,j−1} + F_{1,j}
end for
for i ← 2 to maximal index i in F
    for j ← 2 to maximal index j in F
        D_{i,j} ← F_{i,j} + min {D_{i−1,j}, D_{i,j−1}, D_{i−1,j−1}}
    end for
end for
return D
```

end function

Algorithm A.6 Precomputation of certain minimal subsequence distances

function $D \leftarrow$ precompute_subsequence_distances (A, F, $l_{\min}$, $l_{\max}$, k)

// ***Input:***
// A ... *AAM parameter vectors of all training videos*
// F ... *precomputed frame distances (→ Alg. A.4)*
// l_{min}, l_{max} ... *minimal and maximal length of subsequences that shall be considered*
// k ... *number of best distances to compute for each* (m_1, s_1, l_1, m_2) *parameter combination*
// ***Output:***
// D ... *multi-dimensional matrix of certain minimum subsequence distances:*
// $D^{m_2,i}_{m_1,s_1,l_1}$ ... *the* i*-th minimal distance of the* l_1 *frames long subsequence of the* m_1*-th*
// *video that starts at frame* s_1 *to subsequences of the* m_2*-th video*

$D^{m_2,i}_{m_1,s_1,l_1} \leftarrow +\infty \quad \forall\ m_1, s_1, l_1, m_2, i$
for $m_1 \leftarrow 1$ **to** N
  for $s_1 \leftarrow 1$ **to** $L_{m_1} - l_{\min} + 1$
    for $m_2 \leftarrow 1$ **to** N **with** $m_2 \neq m_1$
      for $s_2 \leftarrow 1$ **to** $L_{m_2} - l_{\min} + 1$
        $F'_{i,j} \leftarrow F^{i+s_1-1,j+s_2-1}_{m_1,m_2} \quad \forall\, i,j:\ 1 \le i \le L_{m_1} - s_1 + 1,\ 1 \le j \le L_{m_2} - s_2 + 1$
        $S \leftarrow$ compute_DTW_distance_matrix(F') //→ *Alg. A.5*
        for $l_1 \leftarrow l_{\min}$ **to** $l_{\max}$ **with** $s_1 + l_1 - 1 \le L_{m_1}$
          for $l_2 \leftarrow \lfloor l_1/f \rceil$ **to** $\lfloor l_1 \cdot f \rceil$ **with** $l_{\min} \le l_2 \le l_{\max} \wedge s_2 + l_2 - 1 \le L_{m_2}$
            $d \leftarrow \frac{S_{l_1,l_2}}{l_1}$
            $i \leftarrow$ minimal i such that $D^{m_2,i}_{m_1,s_1,l_1} \ge d$, if any
            if such an i exists
              for $j \leftarrow k$ **downto** $i+1$
                $D^{m_2,j}_{m_1,s_1,l_1} \leftarrow D^{m_2,j-1}_{m_1,s_1,l_1}$
              end for
              $D^{m_2,i}_{m_1,s_1,l_1} \leftarrow d$
            end if
          end for
        end for
      end for
    end for
  end for
end for
return D

end function

Algorithm A.7 Compute the minimal distances of a sequence to the reference subsequences

function $D^* \leftarrow$ compute_minimal_distances_to_reference_subsequences (R_{success}, R_{failure}, F, m / a)

// ***Input:***
// $R_{success}, R_{failure}$ *... selected reference subsequences →* ***Alg. A.1***
// F *... precomputed frame distances (→* ***Alg. A.4****) or* $\emptyset$ *if invoked during test video classification*
// m *... index of the video sequence in the training videos (in case of* $F \neq \emptyset$ *only)*
// a *... AAM parameter vector sequence of the test video (in case of* $F = \emptyset$ *only)*
// ***Output:***
// D^* *... minimal distances of any subsequences of* a *to all reference subsequences:*
// $D^*_{c,((m_r,s_r,l_r),s_d)}$ *... minimal distance of any subsequence of* a *to reference subsequence*
// (m_r, s_r, l_r) *of class* c *(with discriminativity-score* s_d*)*

for $c \in (success, failure)$
  $l_{\min} \leftarrow l'_r$ with $((m'_r, s'_r, l'_r), s'_d) \leftarrow \min_{l'_r} \{R_c\}$
  for $((m_r, s_r, l_r), s_d) \in R_c$
    $D^*_{c,((m_r,s_r,l_r),s_d)} \leftarrow \infty$
    $L' \leftarrow \begin{cases} L_m & \mid \text{if } F \neq \emptyset \\ L_a & \mid \text{if } F = \emptyset \end{cases}$
    for $s \leftarrow 1$ **to** $L' - \lfloor l_r / f \rceil + 1$
      if $F = \emptyset$
        $F'_{i,j} \leftarrow |A^{i+s_r-1}_{m_r} - a_{i+s-1}| \quad \forall\, i,j:\ 1 \le i \le l_r,\ 1 \le j \le \min\{\lfloor l_r \cdot f \rceil, L_a - s + 1\}$
      else
        $F'_{i,j} \leftarrow F^{i+s_r-1, j+s-1}_{m_r, m} \quad \forall\, i,j:\ 1 \le i \le l_r,\ 1 \le j \le \min\{\lfloor l_r \cdot f \rceil, L_m - s + 1\}$
      end if
      $S \leftarrow$ compute_DTW_distance_matrix(F') //→ *Alg. A.5*
      for $l \leftarrow \lfloor l_r / f \rceil$ **to** $\lfloor l_r \cdot f \rceil$ **with** $1 \le l \le L' - s + 1$
        if $S_{l_r,l} \,/\, l_r < D^*_{c,((m_r,s_r,l_r),s_d)}$
          $D^*_{c,((m_r,s_r,l_r),s_d)} \leftarrow S_{l_r,l} \,/\, l_r$
        end if
      end for
    end for
  end for
  return D^*

end function

Algorithm A.8 Select several reference subsequences for a given video

function $R \leftarrow$ select_reference_subsequences (S, c, m_1, M_2, $l_{\min}$, $l_{\max}$, k, n)

// ***Input:***
// S ... *precomputed subsequence distances (→ Alg. A.6)*
// c ... *class labels of all training videos*
// m_1 ... *index of the training video to select reference subsequences for*
// M_2 ... *set of training video indices for videos to consider in distance evaluations*
// l_{min}, l_{max} ... *minimal and maximal length of subsequences that shall be considered*
// k ... *number of minimal distances to use in discriminativity-score calculation*
// n ... *number of reference subsequences to select*
// ***Output:***
// R ... *selected reference subsequences:*
// $((m_1, s_1, l_1), s) \in R$: *the subsequence in the m_1-th video that starts at frame s_1,*
// *is l_1 frames long, and has a discriminativity-score of s*

$R \leftarrow \emptyset$
for $s_1 \leftarrow 1$ **to** maximal index s_1 in S
  for $l_1 \leftarrow l_{\min}$ **to** $l_{\max}$
    // *Calculate discriminativity-score*
    for $c \in \{success, failure\}$
      $D_c \leftarrow k\_\min \{ S^{m_2,i}_{m_1,s_1,l_1} \mid m_2 \in M_2 \ \wedge \ c_{m_2} = c, \forall i \}$
    end for
    $s \leftarrow \sum_{d \in D_{\bar{c}_{m_1}}} d \ / \ \sum_{d \in D_{c_{m_1}}} d$ // $\bar{c}_{m_1}$ ... *opposite class to* c_{m_1}
    // *Consider this subsequence as reference subsequence*
    if $\|R\| < n \ \vee \ s > s'$ with $((m'_1, s'_1, l'_1), s') \leftarrow \min_{s'} \{ n\_\max_{s'} \{R\} \}$
      if subsequence (m_1, s_1, l_1) does not overlap with any $((m'_1, s'_1, l'_1), s') \in R$ with $s' > s$
        $R \leftarrow$ remove all subsequences from R that overlap with (m_1, s_1, l_1)
        $R \leftarrow R \cup \{ ((m_1, s_1, l_1), s) \}$
      end if
    end if
  end for
end for
return $n\_\max_{s} \{R\}$

end function

Algorithm A.9 Fuse the selected reference subsequences of the single videos

function $(R_{\text{success}}, R_{\text{failure}}) \leftarrow$ fuse_reference_subsequences $(R,\, c,\, v,\, t)$

// ***Input:***
// R ... *reference subsequences of the individual training videos: (→ Alg. A.8)*
// $R_m \in R$ ... *reference subsequences from the m-th training video*
// c ... *class labels of all training videos*
// v ... *number of reference subsequences per training video to select preferably*
// t ... *number of reference subsequences per class in total*
// ***Output:***
// $(R_{success}, R_{failure})$ ... *final reference subsequences for the two classes*

$(S_c, Q_c, R_c) \leftarrow (\emptyset, \emptyset, \emptyset)\ \forall\, c \in \{success, failure\}$
for $m \leftarrow 1$ **to** N
  $S_{c_m} \leftarrow S_{c_m} \cup \underset{s}{v\_\max}\{R_m\}$
  $Q_{c_m} \leftarrow Q_{c_m} \cup R_m \backslash \underset{s}{v\_\max}\{R_m\}$
end for
for $c \in \{success, failure\}$
  if $\|S_c\| \geq t$
    $R_c \leftarrow \underset{s}{t\_\max}\{S_c\}$
  else
    $R_c \leftarrow S_c \cup \underset{s}{(t - \|S_c\|)\_\max}\{Q_c\}$
end for
return $(R_{\text{success}}, R_{\text{failure}})$

end function

Algorithm A.10 Compute the classification scores

function $(d_{\text{success}}, d_{\text{failure}}) \leftarrow$ compute_classification_scores $(D^*,\, t,\, u,\, w)$

// ***Input:***
// D^* ... *minimal distances to reference subsequences (→ Alg. A.7)*
// t ... *number of reference subsequences per class to consider*
// u ... *number of best distances per class to consider*
// w ... *distance weight*
// ***Output:***
// $(d_{success}, d_{failure})$ ... *classification scores for the two classes*

for $c \in \{success, failure\}$
  $\Gamma_c \leftarrow u\_\min \left\{ \underset{s_d}{t\_\max} \{\, D^*_{c,((m_r,s_r,l_r),s_d)} \mid \forall\, ((m_r, s_r, l_r), s_d) \,\} \right\}$
  $d_c \leftarrow \sum_{\gamma \in \Gamma_c} \frac{1}{\gamma^w}$
end for
return $(d_{\text{success}}, d_{\text{failure}})$

end function

Algorithm A.11 Compute the classification bias

function $b \leftarrow$ compute_classification_bias (Y)

$/\!/$ ***Input:***
$/\!/$ Y *... set of classification scores and true classes:*
$/\!/$ $(d_{success}, d_{failure}, c) \in Y$ *... classification scores and true class for one video*
$/\!/$ ***Output:***
$/\!/$ b *... classification bias*

$i \leftarrow 0$
for $(d_{\text{success}}, d_{\text{failure}}, c) \in Y$
  $i \leftarrow i + 1$
  $z_i \leftarrow \frac{d_{\text{success}}}{d_{\text{failure}}}$
end for
$z \leftarrow$ sort z in ascending order
$u_1 \leftarrow \frac{z_1}{1+\epsilon}$ $/\!/ 0 < \epsilon \ll 1$
$u_{i+1} \leftarrow z_n \cdot (1 + \epsilon)$
$u_j \leftarrow \frac{z_j - z_{j-1}}{2} \quad \forall\, j : 2 \leq j \leq i$
$(r_j, \psi_j) \leftarrow$ compute_classification_rate_and_margin $(Y, u_j) \quad \forall\, j : 1 \leq j \leq i+1$ $/\!/ \rightarrow$ *Alg. A.12*
$b \leftarrow u_j$ with $j \leftarrow$ median $\{\arg\max\{r_j \mid 1 \leq j \leq i+1\}\}$
return b

end function

Algorithm A.12 Compute the classification rate and mean classification score margin

function $(r, \psi) \leftarrow$ compute_classification_rate_and_margin (Y, b)

$/\!/$ ***Input:***
$/\!/$ Y *... set of classification scores and true classes:*
$/\!/$ $(d_{success}, d_{failure}, c) \in Y$ *... classification scores and true class for one video*
$/\!/$ b *... classification bias to consider* $\rightarrow$ *Alg. A.11*
$/\!/$ ***Output:***
$/\!/$ (r, ψ) *... classification rate* r *and mean classification score margin* ψ

$\alpha \leftarrow 0$
$(B_{\text{correct}}, B_{\text{wrong}}) \leftarrow (\emptyset, \emptyset)$
for $(d_{\text{success}}, d_{\text{failure}}, c) \in Y$
  if $(d_{\text{success}} \geq b \cdot d_{\text{failure}} \wedge c = \textit{success}) \quad \vee \quad (d_{\text{success}} < b \cdot d_{\text{failure}} \wedge c = \textit{failure})$
    $\alpha \leftarrow \alpha + 1$
    $B_{\text{correct}} \leftarrow B_{\text{correct}} \cup \{(d_{\text{success}}, b \cdot d_{\text{failure}})\}$
  else
    $B_{\text{wrong}} \leftarrow B_{\text{wrong}} \cup \{(d_{\text{success}}, b \cdot d_{\text{failure}})\}$
  end if
end for
$r \leftarrow \frac{\alpha}{\|Y\|}$
$\psi \leftarrow \sum_{(d'_{\text{success}}, d'_{\text{failure}}) \in B_{\text{correct}}} |d'_{\text{success}} - d'_{\text{failure}}| \quad / \quad \sum_{(d'_{\text{success}}, d'_{\text{failure}}) \in B_{\text{wrong}}} |d'_{\text{success}} - d'_{\text{failure}}|$
return (r, ψ)

end function

A.8. Previous Publications

Parts of this dissertation have been published before [305, 308, 306, 307]. In particular, this concerns the following sections and images: Fig. 4.3 and Fig. 4.4 in [304]. Parts of Sec. 3.2, 3.3, 3.5, and 3.6.1, and Fig. 3.2 in [305]. Parts of Sec. 4.5.3, 5.4, and 5.5 in [308]. Parts of Sec. 3.3, 6.1, 6.2, and 6.3.1 in [306]. Parts of Sec. 2.2 (especially Sec. 2.2.3), 3.1, 4.2, 4.3, 4.4, and 4.5.1 in [307]. Christian Lang is the first author and writer of all these publications.

B. Bibliography

[1] R. Abiantun, U. Prabhu, K. Seshadri, J. Heo, and M. Savvides. An Analysis of Facial Shape and Texture for Recognition: A Large Scale Evaluation on FRGC ver2.0. In *Workshop on Applications of Computer Vision*, 2011.

[2] R. B. Adams, Jr. and R. E. Kleck. Perceived Gaze Direction and the Processing of Facial Displays of Emotion. *Psychological Science*, 14(6):644–647, 2003.

[3] J. R. Aiella. A test of equilibrium theory: Visual interaction in relation to orientation, distance and sex of interactants. *Psychonomic Science*, 27:335–336, 1972.

[4] V. Akman. Rethinking context as a social construct. *Journal of Pragmatics*, 32:743–759, 2000.

[5] S. Alirezaee, H. Aghaeinia, K. Faez, and F. Askarig. An Efficient Algorithm for Face Localization. *International Journal of Information Technology*, 12(7), 2006.

[6] A. Amine, S. Ghouzali, and M. Rziza. Face Detection in Still Color Images Using Skin Color Information. In *International Symposium on Communications, Control and Signal Processing*, 2006.

[7] S. Anstis, J. Mayhew, and T. Morley. The Perception of Where a Face or Television 'Portrait' Is Looking. *American Journal of Psychology*, 82(4):474–489, 1969.

[8] M. Argyle. *Social Interaction*. Methuen, 1969.

[9] M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, 1976.

[10] M. Argyle and J. Dean. Eye-Contact, Distance and Affiliation. *Sociometry*, 28(3):289–304, 1965.

[11] M. Argyle and J. A. Graham. The central Europe experiment: Looking at persons and looking at objects. *Journal of Nonverbal Behavior*, 1(1):6–16, 1976.

[12] M. Argyle and R. Ingham. Gaze, Mutual Gaze, and Proximity. *Semiotica*, 6(1):32–49, 1972.

[13] M. B. Arnold. *Emotion & Personality Volume 1: Psychological Aspects*. Columbia University Press, 1960.

[14] C. K. Bainum, K. R. Lounsbury, and H. R. Pollio. The Development of Laughing and Smiling in Nursery School Children. *Child Development*, 55(5):1946–1957, 1984.

[15] S. Baker, I. Matthews, R. Xiao, J. Gross, T. Kanade, and T. Ishikawa. Real-Time Non-Rigid Driver Head Tracking for Driver Mental State Estimation. In *11th World Congress on Intelligent Transportation Systems*, 2004.

[16] V. N. Balasubramanian, J. Ye, and S. Panchanathan. Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[17] S. Baluja and D. Pomerleau. Non-Intrusive Gaze Tracking Using Artificial Neural Networks. Technical report, Carnegie Mellon University, 1994.

[18] P. Barkhuysen, E. Krahmer, and M. Swerts. Problem Detection in Human-Machine Interactions based on Facial Expressions of Users. *Speech communication*, 45(3):343–359, 2005.

[19] S. Baron-Cohen. How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Cahiers De Psychologie Cognitive*, 13(5):513–552, 1994.

[20] S. Baron-Cohen. *Mindblindness - An Essay on Autism and Theory of Mind*. MIT Press, 1996.

[21] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully Automatic Facial Action Recognition in Spontaneous Behavior. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 223–230, 2006.

[22] M. S. Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California, San Diego, 1998.

[23] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Conference on Computer Vision and Pattern Recognition*, pages 568–573, 2005.

[24] A. Battocchi, F. Pianesi, and D. Goren-Bar. The Properties of DaFEx, a Database of Kinetic Facial Expressions. In *Proceedings of the first International Conference on Affective Computing and Intelligent Interaction*, volume 3784 of *Lecture Notes in Computer Science*, pages 558–565, Beijing, China, October 2005.

[25] J. B. Bavelas, A. Black, C. R. Lemery, and J. Mullett. I Show How You Feel: Motor Mimicry as a Communicative Act. *Journal of Personality and Social Psychology*, 50(2):322–329, 1986.

[26] J. B. Bavelas and N. Chovil. *The Psychology of Facial Expression*, chapter Faces in Dialogue, pages 334–346. Cambridge University Press, 1997.

[27] J. B. Bavelas and N. Chovil. Visible Acts of Meaning: An Integrated Message Model of Language in Face-to-Face Dialogue. *Journal of Language and Social Psychology*, 19(2):163–194, 2000.

[28] M. Belkin and P. Niyogi. Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[29] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[30] D. J. Beymer. Face Recognition Under Varying Pose. In *Conference on Computer Vision and Pattern Recognition*, pages 756–761, 1994.

[31] S. K. Bhatia, V. Lakshminarayanan, A. Samal, and G. V. Welland. Human Face Perception in Degraded Images. *Journal of Visual Communication and Image Representation*, 6(3):280–295, 1995.

[32] M. Bindemann, A. M. Burton, and S. R. H. Langton. How do eye gaze and facial expression interact? *Visual Cognition*, 16(6):708–733, 2007.

[33] R. Birdwhistell. *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press, 1970.

[34] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

[35] A. Bjõrck. *Numerical Methods for Least Squares Problems*. Society for Industrial Mathematics, 1996.

[36] M. J. Black and A. D. Jepson. A Probabilistic Framework for Matching Temporal Trajectories: CONDENSATION-Based Recognition of Gestures and Expressions. In *5th European Conference on Computer Vision*, volume 1, pages 909–924, 1998.

[37] F. L. Bookstein. Landmark methods for forms without landmarks: localizing group differences in outline shape. *Medical Image Analysis*, 1(3):225âĂŞ–244, 1997.

[38] J. D. Boucher and G. E. Carlson. Recognition of Facial Expression in Three Cultures. *Journal of Cross-Cultural Psychology*, 11:263–280, 1980.

[39] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[40] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 1st edition, 2008.

[41] G. Breed. The Effect of Intimacy: Reciprocity or Retreat? *British Journal of Social and Clinical Psychology*, 11(2):135–142, 1972.

[42] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[43] V. J. Brightman, A. L. Segal, P. Werther, and J. Steiner. Ethologic study of facial expressions in response to taste stimuli. *Journal of Dental Research*, 54:L141 (Abstract), 1975.

[44] V. J. Brightman, A. L. Segal, P. Werther, and J. Steiner. Facial expression and hedonic response to face stimuli. *Journal of Dental Research*, 56:B161 (Abstract), 1977.

[45] V. Bruce. Influences of familiarity on the processing of faces. *Perception*, 15(4):387–397, 1986.

[46] V. Bruce, P. R. Green, and M. A. Georgeson. *Visual Perception: Physiology, Psychology and Ecology*. Psychology Press, fourth edition, 2003.

[47] V. Bruce, Z. Henderson, K. Greenwood, P. J. B. Hancock, A. M. Burton, and P. Miller. Verification of face identities from images captured on video. *Journal of Experimental Psychology*, 5(4):339–360, 1999.

[48] V. Bruce and S. Langton. The use of pigmentation and shading information in recognising the sex and identities of faces. *Perception*, 23(7):803–822, 1994.

[49] V. Bruce and A. Young. Understanding Face Recognition. *British Journal of Psychology*, 77(3):305–327, 1986.

[50] R. Bruyer and G. Crispeels. Expertise in person recognition. *Bulletin of the Psychonomic Society*, 30(6):501–504, 1992.

[51] J. Buenaposada and L. Baumela. Variations of Grey World for Face Tracking. *Image Processing & Communications*, 7(3–4):51–61, 2001.

[52] J. Buenaposada and L. Baumela. Real-time tracking and estimation of plane pose. In *International Conference on Pattern Recognition*, volume 2, pages 697–700, 2002.

[53] J. M. Buenaposada, E. Muñoz, and L. Baumela. Recognising facial expressions in video sequences. *Pattern Analysis & Applications*, 11(1):101–116, 2008.

[54] C. M. Bukach, I. Gauthier, and M. J. Tarr. Beyond faces and modularity: the power of an expertise framework. *Trends in Cognitive Sciences*, 10(4):159–166, 2006.

[55] A. M. Burton, S. Wilson, M. Cowan, and V. Bruce. Face Recognition in Poor-Quality Video: Evidence from Security Surveillance. *Psychological Science*, 10(3):243–248, 1999.

[56] L. E. Bush. Individual differences multidimensional scaling of adjectives denoting feelings. *Journal of Personality and Social Psychology*, 25(1):50–57, 1973.

[57] G. Butterworth. Towards a Mechanism of Joint Visual Attention in Human Infancy. *International Journal of Behavioral Development*, 3(3):253–272, 1980.

[58] J. Cai and A. Goshtasby. Detecting human faces in color images. *Image and Vision Computing*, 18(1):63–75, 1999.

[59] A. J. Calder and A. W. Young. Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6:641–651, 2005.

[60] J. Canny. A Computational Approach to Edge Detection. *Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[61] S. Carey and R. Diomand. From piecemeal to configurational representation of faces. *Science*, 21:312–314, 1977.

[62] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *8th International Conference on Multimodal Interfaces*, pages 146–154, 2006.

[63] J. M. Carroll and J. A. Russell. Facial Expressions in Hollywood's Portrayal of Emotion. *Journal of Personality and Social Psychology*, 72(1):164–176, 1997.

[64] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández. ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, 18(2):130–140, 2007.

[65] M. Castrillón, O. Déniz, and M. Hernández. The ENCARA System for Face Detection and Normalization. *Lecture Notes in Computer Science*, 2652:176–âĂŞ183, 2003.

[66] Centre for Vision, Speech and Signal Processing, University of Surrey. Recognition And Vision Library (RAVL). `http://www.ee.surrey.ac.uk/CVSSP/Ravl/RavlDoc/share/doc/RAVL/Auto/Basic/Tree/Ravl.html`. visited on 2012-06-18.

[67] L. Cerrato and M. Skhiri. A method for the analysis and measurement of communicative head movements in human dialogues. In *International Conference on Audio-Visual Speech Processing*, pages 251–256, 2003.

[68] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *Conference on Computer Vision and Pattern Recognition*, pages 2567–2573, 2010.

[69] D. Chai and K. N. Ngan. Locating Facial Region of a Head-and-Shoulders Color Imageinternational conference on face & gesture recognition. In *International Conference on Face and Gesture Recognition*, pages 124–129, 1998.

[70] C.-C. Chang and C.-J. Lin. LIBSVM - A Library for Support Vector Machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. visited on 2012-06-18.

[71] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[72] W. A. Chaovalitwongse and P. M. Pardalos. On the time series support vector machine using dynamic time warping kernel for brain activity classification. *Cybernetics and Systems Analysis*, 14(1):125–138, 2008.

[73] A. J. Chapman. Eye Contact, Physical Proximity and Laughter: A Re-Examination of the Equilibrium Model of Social Intimacy. *Social Behavior and Personality*, 3(2):143–156, 1975.

[74] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and Machine Recognition of Faces: A Survey. *Proceedings of the IEEE*, 83(5):705–741, 1995.

[75] L. Chen and R. Ng. On The Marriage of Lp-norms and Edit Distance. In *Proceedings of the Thirtieth international conference on Very large data bases*, volume 30, pages 792–803, 2004.

[76] L. S.-H. Chen. *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction.* PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[77] D. Chetverikov and A. Lerch. Multiresolution face detection. In *Theoretical Foundations of Computer Vision*, volume 69, pages 130–140, 1993.

[78] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic Discovery of Time Series Motifs. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pages 493–498, 2003.

[79] N. Chovil. Discourse-Oriented Facial Displays in Conversation. *Research on Language and Social Interaction*, 25:163–194, 1991.

[80] N. Chovil. Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15(3):141–154, 1991.

[81] V. P. Clark, K. Keil, J. M. Maisog, S. Courtney, L. G. Ungerleider, and J. V. Haxby. Functional Magnetic Resonance Imaging of Human Visual Cortex during Face Matching: A Comparison with Positron Emission Tomography. *NeuroImage*, 4(1):1–15, 1996.

[82] M. G. Cline. The Perception of Where a Person Is Looking. *The American Journal of Psychology*, 80(1):41–50, 1967.

[83] J. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *International Conference on Systems, Man and Cybernetics*, pages 610–616, 2004.

[84] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression. In *International Conference on Face and Gesture Recognition*, pages 396–401, 1998.

[85] R. M. Cooper. *The Effects of Eye Gaze and Emotional Facial Expression on the Allocation of Visual Attention.* PhD thesis, University of Stirling, Department of Psychology, 2006.

[86] T. Cootes, D. Cooper, C. Taylor, and J. Graham. Active Shape Models — Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, Januar 1995.

[87] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In H. Burkhardt and B. Neumann, editors, *Proceedings European Conference on Computer Vision*, volume 2, pages 484–498. Springer, 1998.

[88] T. Cootes and C. Taylor. An Algorithm for Tuning an Active Appearance Model to New Data. In *Proceedings of the British Machine Vision Conference*, volume 3, pages 919–928, 2006.

[89] T. F. Cootes, G. Edwards, and C. J. Taylor. Comparing Active Shape Models with Active Appearance Models. In *British Machine Vision Conference*, volume 1, pages 173–182, 1999.

[90] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. *Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[91] T. F. Cootes and C. J. Taylor. Active Shape Models - 'Smart Snakes'. In *British Machine Vision Conference*, pages 266–275, 1992.

[92] T. F. Cootes and C. J. Taylor. Locating faces using statistical feature detectors. In *International Conference on Automatic Face and Gesture Recognition*, 204–209, 1996.

[93] T. F. Cootes, K. Walker, and C. J. Taylor. View-based active appearance models. In *International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.

[94] M. D. Cordea, E. M. Petriu, N. D. Georganas, D. C. Petriu, and T. E. Whalen. Real-Time 2(1/2)-D Head Pose Recovery for Model-Based Video-Coding. *Transactions on Instrumentation and Measurement*, 50(4):1007–1013, 2001.

[95] L. M. Coutts and F. W. Schneider. Visual behavior in an unfocused interaction as a function of sex and distance. *Journal of Experimental Social Psychology*, 11(1):64–77, 1975.

[96] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. ŕFEELTRACE: An instrument for recording perceived emotion in real time. In *ISCA Workshop on Speech and Emotion*, pages 19–24, 2000.

[97] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion Recognition in Human-Computer Interaction. *Signal Processing Magazine*, 18(1):32–80, 2001.

[98] I. Craw, H. Ellis, and J. R. Lishman. Automatic extraction of face-features. *Pattern Recognition Letters*, 5(2):183–187, 1987.

[99] I. Craw, D. Tock, and A. Bennett. Finding Face Features. In *European Conference on Computer Vision*, volume 588/1992 of *Lecture Notes in Computer Science*, pages 92–96, 1992.

[100] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[101] D. Cristinacce and T. Cootes. Feature Detection and Tracking with Constrained Local Models. In *Proceedings of the British Machine Vision Conference*, volume 3, pages 929–938, 2006.

[102] J. L. Crowley and F. Berard. Multi-Modal Tracking of Faces for Video Communications. In *Conference on Computer Vision and Pattern Recognition*, pages 640–645, 1997.

[103] M. N. Dailey and G. W. Cottrell. PCA = Gabor for Expression Recognition. Technical Report CS–629. Technical report, Computer Science and Engineering, University of California, San Diego, 1999.

[104] K. Dautenhahn. Faciliating Social Interaction with Robot Companions. Talk at Bielefeld University, May 5th, 2011.

[105] J. Davidoff and N. Donnelly. Object superiority: A comparison of complete and part probes. *Acta Psychologica*, 73(3):225–243, 1990.

[106] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, 2. edition, 2000.

[107] I. de Kok and D. Heylen. The MultiLis Corpus - Dealing with Individual Differences in Nonverbal Listening Behavior. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues - Third COST 2102 International Training School*, pages 362–375, 2010.

[108] D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 231–238, 1996.

[109] D. DeCarlo, M. Stone, C. Revilla, and J. J. Venditti. Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38, 2004.

[110] N. Degtyarev and O. Seredin. Comparative Testing of Face Detection Algorithms. In *International Conference on Image and Signal Processing*, volume 6134/2010 of *Lecture Notes in Computer Science*, pages 200–209, 2010.

[111] A. Demiriz, K. P. Bennett, and J. Shawe-Taylor. Linear Programming Boosting via Column Generation. *Machine Learning*, 46(1–3):225–254, 2002.

[112] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[113] R. Diamond and S. Carey. Why Faces Are and Are Not Special: An Effect of Expertise. *Journal of Experimental Psychology: General*, 115(2):107–117, 1986.

[114] D. Dicks. *Experimental study of natural conversation.* PhD thesis, Imperial College London, 1972.

[115] H. Ding, G. Trajcevski, P. Scheuermann, and E. Wang, Xiaoyue Keogh. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. *Proceedings of the VLDB Endowment (PVLDB)*, 1(2):1542–1552, 2008.

[116] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974âĂŞ–989, 1999.

[117] N. Donnelly and J. Davidoff. The Mental Representations of Faces and Houses: Issues Concerning Parts and Wholes. *Visual Cognition*, 6(3/4):319–343, 1999.

[118] G. Doretto and S. Soatto. Dynamic Shape and Appearance Models. *Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2006–2019, 2006.

[119] F. Dornaika and F. Davoine. Head and Facial Animation Tracking using Appearance-Adaptive Models and Particle Filters. In *Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[120] A. Doucet, N. De Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[121] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1–2):33–60, 2003.

[122] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 488–500, 2007.

[123] N. Dresser. *Multicultural Manners: Essential Rules of Etiquette for the 21st Century*. Wiley & Sons, 2005.

[124] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support Vector Regression Machines. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 155–161, 1996.

[125] L. Ducci, L. Arcuri, T. W/ Georgis, and T. Sineshaw. Emotion Recognition in Ethiopia: The Effect of Familiarity with Western Culture on Accuracy of Recognition. *Journal of Cross-Cultural Psychology*, 13:340–351, 1982.

[126] R. O. Duda and P. E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[127] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2. edition, 2000.

[128] S. Duncan and D. W. Fiske. *Face to Face Interaction*. Lawrence Erlbaum Associates, 1977.

[129] G. Edwards, T. Cootes, and C. Taylor. Face Recognition Using Active Appearance Models. In H. Burkhardt and B. Neumann, editors, *Proceeding of the European Conference on Computer Vision*, volume 2, pages 581–695. Springer, 1998.

[130] G. J. Edwards, A. Lanitis, C. J. Taylor, and T. Cootes. Statistical models of face images — improving specificity. *Image and Vision Computing*, 16(3):203–211, 1998.

[131] J. S. Efran. Looking for approval: Effects on visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology*, 10(1):21–25, 1968.

[132] N. Eilan, C. Hoerl, T. McCormack, and J. Roessler, editors. *Joint Attention: Communication and Other Minds: Issues in Philosophy and Psychology (Consciousness and Self-Consciousness)*. Oxford University Press, 2005.

[133] P. Ekman. Universals and Cultural Differences in Facial Expressions of Emotion. *Nebraska Symposium on Motivation*, 19:207–283, 1971.

[134] P. Ekman. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3 & 4):169–200, 1992.

[135] P. Ekman. Facial Expression of Emotion. *American Psychologist*, 48(4):384–392, 1993.

[136] P. Ekman. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994.

[137] P. Ekman. Should We Call It Expression Or Communication? *Innovation*, 10(4):333–344, 1997.

[138] P. Ekman, R. J. Davidson, and W. V. Friesen. The Duchenne Smile: Emotional Expression and Brain Physiology II. *Journal of Personality and Social Psychology*, 58(2):342–353, 1990.

[139] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[140] P. Ekman and W. V. Friesen. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, 1(1):49–98, 1969.

[141] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, 1987.

[142] P. Ekman, J. C. Hager, and W. V. Friesen. The Symmetry of Emotional and Deliberate Facial Actions. *Psychophysiology*, 18(2):101–106, 1981.

[143] P. Ekman, E. R. Sorenson, and W. V. Friesen. Pan-Cultural Elements in Facial Displays of Emotion. *Science*, 164:86–88, 1969.

[144] R. El Kaliouby and P. Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. *Transactions on Robotics*, 23(5):991–1000, 2007.

[145] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000.

[146] D. Erickson, O. Fujimura, and B. Pardo. Articulatory correlates of prosodic control: Emotion versus emphasis. *Language and Speech - Special Issue on Prosody and Conversation*, 41(3–4):395–413, 1998.

[147] R. Exline, D. Gray, and D. Schuette. Visual behavior in a dyad as affected by interview content and sex of respondent. *Journal of Personality and Social Psychology*, 1(3):201–209, 1965.

[148] R. V. Exline. Explorations in the process of person perception: visual interaction in relation to competition, sex, and need for affiliation. *Journal of Personality*, 31(1):1–20, 1963.

[149] R. V. Exline. Visual interaction: The glances of power and preference. *Nebraska Symposium on Motivation*, 19:163–206, 1972.

[150] R. V. Exline, I. Gottheil, A. Paredes, and D. Winklemeier. Gaze Direction as a Factor in Judgment of Non-Verbal Expressions of Affect. In *76th Annual Conference of the American Psychological Association*, pages 415–416, 1968.

[151] R. V. Exline, J. Thibaut, C. B. Hickey, and P. Gumpert. *Studies in Machiavellianism*, chapter Visual interaction in relation to machiavellianism and an unethical act, pages 53–76. Academic Press, 1970.

[152] R. V. Exline and L. C. Winters. *Affect, cognition and personality*, chapter Affective relations and mutual glances in dyads. Springer, New York, 1965.

[153] R. V. Exline and L. C. Winters. *Affect, cognition, and personality*, chapter Affective relations and mutual glances in dyads, pages 319–350. Tavistock, 1966.

[154] FaceTracker. A C/C++ API for real time generic non-rigid face alignment and tracking. `http://web.mac.com/jsaragih/FaceTracker/FaceTracker.html`. visited on 2012-06-26.

[155] G. Fanelli, J. Gal, and L. Van Gool. Real Time Head Pose Estimation with Random Regression Forests. In *Conference on Computer Vision and Pattern Recognition*, pages 617–624, 2011.

[156] M. J. Farah. *Behavioral Neurology and Neuropsychology*, chapter Prosopagnosia, pages 239–241. McGraw-Hill, second edition, 2003.

[157] M. J. Farah and K. D. Wilson. What Is "Special" About Face Perception. *Psychological Review*, 105(3):482–498, 1998.

[158] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003.

[159] Z. Fei and Q. Qiang. Face Detection Based on Rectangular Knowledge Rule and Face Structure. In *International Conference on Information Science and Engineering*, pages 1235–1239, 2009.

[160] W. A. Fellenz, J. G. Taylor, N. Tsapatsoulis, and S. Kollias. Comparing Template-based, Feature-based and Supervised Classification of Facial Expressions from Static Images. In *Proceedings of Circuits, Systems, Communications and Computers*, pages 5331–5336, 1999.

[161] J.-M. Fernández-Dols and M.-A. Ruiz-Belda. Are Smiles a Sign of Happiness? Gold Medal Winners at the Olympic Games. *Journal of Personality and Social Psychology*, 69(6):1113–1119, 1995.

[162] J. M. Fernández-Dols and M.-A. Ruiz-Belda. *The Psychology of Facial Expression*, chapter Spontaneous facial behavior during intense emotional episodes: Artistic truth and optical truth, pages 255–274. Cambridge University Press, 1997.

[163] G. A. Fink. *Mustererkennung mit Markov-Modellen: Theorie - Praxis - Anwendungsgebiete*. Leitfäden der Informatik. Vieweg+Teubner Verlag, 2003.

[164] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179–188, 1936.

[165] A. Floratou, S. Tata, and J. M. Patel. Efficient and Accurate Discovery of Patterns in Sequence Data Sets. *Transactions on Knowledge and Data Engineering*, 23(8):1154–1168, 2011.

[166] N. Fragopanagos and J. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[167] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[168] A. J. Fridlund. Sociality of Solitary Smiling: Potentiation by an Implicit Audience. *Journal of Personality and Social Psychology*, 60(2):229–240, 1991.

[169] A. J. Fridlund. *Human facial expression: An evolutionary view.* Academic Press, San Diego, CA, 1994.

[170] A. J. Fridlund, J. P. Sabini, L. E. Hedlund, J. A. Schaut, J. I. Shenker, and M. J. Knauer. Audience effects on solitary faces during imagery: Displaying to the people in your head. *Journal of Nonverbal Behavior*, 14(2):113–137, 1990.

[171] J. Friedman, T. Hastie, and R. Tibshirani. Additive Logistic Regression: a Statitical View of Boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[172] C. K. Friesen and A. Kingstone. The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, 5(3):490–495, 1998.

[173] C. K. Friesen, C. Moore, and A. Kingstone. Does gaze direction really trigger a reflexive shift of spatial attention? *Brain and Cognition*, 57(1):66–69, 2005.

[174] N. H. Frijda. An understanding of facial expression of emotion. *Acta Psychologica*, 9:294–362, 1953.

[175] N. H. Frijda and A. Tcherkassof. *The Psychology of Facial Expression*, chapter Facial expressions as modes of action readiness, pages 78–102. Cambridge University Press, 1997.

[176] R. Fry and G. F. Smith. The effects of feedback and eye contact on performance of a digit-coding task. *Journal of Social Psychology*, 96(1):145–146, 1975.

[177] B. V. Funt, K. Barnard, and L. Martin. Is Machine Colour Constancy Good Enough? In *European Conference on Computer Vision*, volume I, pages 445–459, 1998.

[178] R. E. Galper. Recognition of faces in photographic negative. *Psychonomic Science*, 19(4):207–208, 1970.

[179] I. Gauthier. *Attention and Performance XXI: Processes of Change in Brain and Cognitive Development*, chapter Constraints on the acquisition of specialization for face processing, pages 275–284. Oxford University Press, first edition edition, 2006.

[180] I. Gauthier and C. Bukach. Should we reject the expertise hypothesis? *Cognition*, 103(2):322–330, 2007.

[181] I. Gauthier, P. Skudlarski, J. C. Gore, and A. W. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2):191–197, 2000.

[182] I. Gauthier and M. J. Tarr. Becoming a "Greeble" Expert: Exploring Mechanisms for Face Recognition. *Vision Research*, 37(12):1673–1682, 1997.

[183] I. Gauthier and M. J. Tarr. Unraveling Mechanisms for Expert Object Recognition: Bridging Brain Activity and Behavior. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):431–446, 2002.

[184] I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2(6):568–573, 1999.

[185] I. Gauthier, P. Williams, M. J. Tarr, and J. Tanaka. Training 'greeble' experts: a framework for studying expert object recognition processes. *Vision Research*, 38(15–16):2401–2428, 1998.

[186] A. Gee and R. Cipolla. Determining the Gaze of Faces in Images. *Image and Vision Computing*, 14(2):639–648, 1994.

[187] P. Geurts. Pattern Extraction for Time Series Classification. In *European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 115–127, 2001.

[188] A. N. Gilbert, A. J. Fridlund, and J. Sabini. Hedonic and social determinants of facial displays to odors. *Chemical Senses*, 12(2):355–363, 1987.

[189] E. Goffman. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, 1963.

[190] J. Gonzalez-Mora, F. De la Torre, R. Murthi, N. Guil, and E. L. Zapata. Bilinear Active Appearance Models. Workshop on Non-rigid Registration and Tracking through Learning, October 2007.

[191] M. H. Goodwin and C. Goodwin. Gesture and coparticipation in the activity of searching for a word. *Semiotica*, 62(1–2):51–76, 1986.

[192] N. Gourier, D. Hall, and J. L. Crowley. Estimating Face orientation from Robust Detection of Salient Facial Structures. In *ICPR Workshop on Visual Observation of Deictic Gestures*, pages 17–25, 2004.

[193] J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

[194] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petaja. Multi-Modal System for Locating Heads and Faces. In *International Conference on Automatic Face and Gesture Recognition*, pages 88–93, 1996.

[195] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang. Visual Prosody: Facial Movements Accompanying Speech. In *International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 2002.

[196] K. Grammer, W. Schiefenhövel, M. Schleidt, B. Lorenz, and I. Eibl-Eibesfeldt. Patterns on the Face: The Eyebrow Flash in Crosscultural Comparison. *Ethology*, 77(4):279–299, 1988.

[197] R. F. Green and M. R. Goldfried. On the bipolarity of semantic space. *Psychological Monographs: General & Applied*, 79(6, Whole No. 599):31, 1965.

[198] R. Gross, I. Matthews, and S. Baker. Generic vs. Person Specific Active Appearance Models. *Image and Vision Computing*, 23(12):1080–1093, 2005.

[199] R. Gross, I. Matthews, and S. Baker. Active Appearance Models with Occlusion. *Image and Vision Computing*, 24(6):593–604, 2006.

[200] C. Guerra. *Contribuciones al Seguimiento Visual Precategórico*. PhD thesis, Universidad de Las Palmas de Gran Canaria, 2002.

[201] Z. Gui and C. Zhang. 3D Head Pose Estimation Using Non-rigid Structure-from-motion and Point Correspondence. In *Region 10 Conference*, pages 1–3, 2006.

[202] H. Gunes and M. Pantic. Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners. In *International Conference on Intelligent Virtual Agents*, pages 371–377, 2010.

[203] H. Gunes and M. Piccardi. A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior. In *International Conference on Pattern Recognition*, pages 1148–1153, 2006.

[204] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. BIRON - The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini, and M. Haegele, editors, *Proceedings of the International Workshop on Advances in Service Robotics*, pages 27–32, Stuttgart, May 2004. Fraunhofer IRB Verlag.

[205] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1–2):35–46, 1983.

[206] M. A. K. Halliday and R. Hasan. *Cohesion in English.* Addison-Wesley Longman, Amsterdam, 1976.

[207] M. Hanheide, S. Wrede, C. Lang, and G. Sagerer. Who am I talking with? A Face Memory for Social Robots. In *International Conference on Robotics and Automation*, pages 3360–3365, 2008.

[208] S. J. Hanson and Y. O. Halchenko. Brain Reading Using Full Brain Support Vector Machines for Object Recognition: There Is No "Face" Identification Area. *Neural Computation*, 20(2):486–503, 2008.

[209] A. Haro, A. Schodl, and I. A. Essa. Head Tracking Using a Textured Polygonal Model. In *Workshop on Perceptual User Interfaces*, 1998.

[210] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539):2425–2430, 2001.

[211] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–233, 2000.

[212] J. V. Haxby, B. Horwitz, L. G. Ungerleider, J. M. Maisog, P. Pietrini, and C. L. Grady. The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. *Journal of Neuroscience*, 14(11):6336–6353, 1994.

[213] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 2nd edition, 1998.

[214] D. Heylen. Challenges Ahead: Head movements and other social acts in conversations. In *Joint Symposium on Virtual Social Agents*, pages 45–52, 2005.

[215] D. Heylen. Facial displays, emotional expressions and conversational acts. In *Cybernetics and Systems*, pages 18–21, 2006.

[216] D. Heylen. Head Gestures, Gaze and the Principles of Conversational Structure. *International Journal of Humanoid Robotics*, 3(3):1–27, 2006.

[217] D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud. Searching for Prototypical Facial Feedback Signals. In *International Conference on Intelligent Virtual Agents*, pages 147–153, 2007.

[218] D. Heylen, A. Nijholt, and D. Reidsma. Determining what people feel and think when interacting with humans and machines - Notes on corpus collection and annotation. In *First California Conference on Recent Advances in Engineering Mechanics*, pages 1–6, 2006.

[219] S. Hinte and M. Lohse. The Function of Off-Gaze in Human-Robot Interaction. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 87–92, 2011.

[220] E. Hjelmås and I. Farup. A Comparison of Face/Non-face Classiffiers. In *International Conference on Audio- and Video-Based Person Authentication*, volume 2091/2001 of *Lecture Notes in Computer Science*, pages 65–70, 2001.

[221] E. Hjelmås and B. K. Low. Face Detection: A Survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[222] N. Hofemann. *Videobasierte Handlungserkennung für die natürliche Mensch-Maschine-Interaktion*. PhD thesis, Bielefeld University, Faculty of Technology, 2007.

[223] G. J. Hole. Configurational factors in the perception of unfamiliar faces. *Perception*, 23(1):65–74, 1994.

[224] H. Hong, H. Neven, and C. von der Malsburg. Online facial expression recognition based on personalized galleries. In *International Conference on Automatic Face and Gesture Recognition*, pages 354–359, 1998.

[225] B. M. Hood, J. D. Willen, and J. Driver. Adult's Eyes Trigger Shifts of Visual Attention in Human Infants. *Psychological Science*, 9(2):131–134, 1998.

[226] T. Horprasert, Y. Yacoob, and L. S. Davis. Computing 3-D head orientation from a monocular image sequence. In *International Conference on Automatic Face and Gesture Recognition*, pages 242–247, 1996.

[227] P. V. C. Hough. Machine Analysis of Bubble Chamber Pictures. In *International Conference on High Energy Accelerators and Instrumentation*, pages 554–556, 1959.

[228] P. O. Hoyer. Non-negative sparse coding. In *Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.

[229] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face Detection in Color Images. *Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, 2002.

[230] C. Hu, J. Xiao, I. Matthews, S. Baker, J. Cohn, and T. Kanade. Fitting a Single Active Appearance Model Simultaneously to Multiple Images. In *British Machine Vision Conference*, pages 437–446, 2004.

[231] C. Huang, H. Ai, T. Yamashita, S. Lao, and M. Kawade. Incremental Learning of Boosted Face Detector. In *International Conference on Computer Vision*, pages 1–8, 2007.

[232] C.-L. Huang and Y.-M. Huang. Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. *Journal of Visual Communication and Image Representation*, 8(3):278–290, 1997.

[233] J. Huang, S. Gutta, and H. Wechsler. Detection of Human Faces Using Decision Trees. In *International Conference on Automatic Face and Gesture Recognition*, pages 248–252, 1996.

[234] Y.-S. Huang. A Robust Multiple Oriented-Template Face Detector. *IAENG International Journal of Computer Science*, 35(3):421–426, 2008.

[235] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, second edition, 2008.

[236] V. Hugot. Eye Gaze Analysis in Human-Human Interactions. Master's thesis, KTH Royal Institute of Technology, School of Computer Science and Communication, Stockholm, Sweden, 2007.

[237] F. A. Huppert and J. E. Whittington. Evidence for the independence of positive and negative well-being: Implications for quality of life assessment. *British Journal of Health Psychology*, 8:107–122, 2003.

[238] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[239] iCub. An open source cognitive humanoid robotic platform. `http://www.icub.org/`. visited on 2012-06-26.

[240] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, K. C. Mailis, Theofilos P.and Karpouzis, and S. D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4):423–435, 2005.

[241] M. Isard and A. Blake. CONDENSATION — Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[242] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive Driver Gaze Tracking with Active Appearance Models. In *11th World Congress on Intelligent Transportation Systems*, 2004.

[243] R. J. Itier. Face Recognition Memory and Configural Processing: A Developmental ERP Study using Upright, Inverted, and Contrast-Reversed Faces. *Journal of Cognitive Neuroscience*, 16(3):487–502, 2004.

[244] P. Ivan. Active Appearance Models for Gaze Estimation. Master's thesis, Vrije Universiteit Amsterdam, Faculty of Sciences, Business Mathematics & Informatics, 2007.

[245] Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, and K. Shirai. Analysis of head movements and its role in spoken dialogue. In *Fourth International Conference on Spoken Language*, volume 4, pages 2167–2170, 1996.

[246] C. E. Izard. *The Face of Emotion*. Appleton-Century-Crofts, 1971.

[247] C. E. Izard. *Human Emotions*. Plenum Press, 1977.

[248] C. E. Izard. Innate and Universal Facial Expressions: Evidence From Developmental and Cross-Cultural Research. *Psychological Bulletin*, 115(2):288–299, 1994.

[249] C. E. Izard. The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation, and Regulation. *Emotion Review*, 2(4):363–370, 2010.

[250] E. Jakobs, A. S. R. Manstead, and A. H. Fischer. Social context effects on facial activity in a negative emotional setting. *Emotion*, 1(1):51–69, 2001.

[251] J.-S. Jang and J.-H. Kim. Fast and Robust Face Detection Using Evolutionary Pruning. *Transactions on Evolutionary Computation*, 12(5):562–571, 2008.

[252] S.-H. Jeng, H. Y. M. Liao, C. C. Han, M. Y. Chern, and Y. T. Liu. Facial feature detection using geometrical face model: An efficient approach. *Pattern Recognition*, 31(3):273–282, 1998.

[253] X. Ji, J. Bailey, and G. Dong. Mining Minimal Distinguishing Subsequence Patterns with Gap Constraints. *Knowledge and Information Systems*, 11(3):259–286, 2007.

[254] Z. Jin, Z. Lou, J. Yang, and Q. Sun. Face Detection Using Template Matching and Skin Color Information. In *International Conference on Intelligent Computing*, pages 636–645, 2005.

[255] Z. Jin, Z. Lou, J. Yang, and Q. Sun. Face detection using template matching and skin-color information. *Neurocomputing*, 70(4–6):794–800, 2007.

[256] M. H. Johnson. Subcortical face processing. *Nature Reviews Neuroscience*, 6:766–774, 2005.

[257] M. H. Johnson, S. Dziurawiec, H. Ellis, and J. Morton. Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1–2):1–19, 1991.

[258] M. H. Johnson and J. Morton. *Biology and Cognitive Development: The Case of Face Recognition*. Blackwell, first edition, 1991.

[259] A. Johnston, H. Hill, and N. Carman. Recognising faces: effects of lighting direction, inversion, and brightness reversal. *Perception*, 21(3):365–375, 1992.

[260] M. Jones and P. Viola. Fast Multi-view Face Detection. Technical Report TR2003-96, Mitsubishi Electric Research Laboratories, 2003.

[261] M. J. Jones and J. M. Rehg. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[262] J. Jonides. *Attention and Performance*, volume IX, chapter Voluntary versus Automatic Control over the Mind's Eye's Movement, pages 187–203. Psychology Press, 1984.

[263] L. P. Kaelbling, M. L. Littman, and A. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[264] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal Of Basic Engineering*, 82(Series D):35–45, 1960.

[265] T. Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces*. PhD thesis, Department of Information Science, Kyoto University, 1973.

[266] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive Database for Facial Expression Analysis. In *Proceedings of the International Conference on automatic Face and Gesture Recognition*, pages 46–53, 2000.

[267] H.-B. Kang. Face Detection with an Adaptive Skin Color Segmentation and Eye Features. In *Intelligent Computing in Signal Processing and Pattern Recognition*, volume 345/2006 of *Lecture Notes in Control and Information Sciences*, pages 852–857, 2006.

[268] N. Kanwisher. Domain specificity in face perception. *Nature Neuroscience*, 3(8):759–763, 2000.

[269] N. Kanwisher. What's in a Face? *Science*, 311(5761):617–618, 2006.

[270] N. Kanwisher, J. McDermott, and M. M. Chun. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

[271] N. Kanwisher, D. Stanley, and A. Harris. The fusiform face area is selective for faces not animals. *Neuroreport*, 10(1):183–187, 1999.

[272] N. Kanwisher and G. Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London B*, 361:2109–2128, 2006.

[273] A. Kapoor and R. W. Picard. Multimodal Affect Recognition in Learning Environments. In *13th Annual ACM international Conference on Multimedia*, pages 677–682, 2005.

[274] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[275] D. G. Kendall. A Survey of the Statistical Theory of Shape. *Statistical Science*, 4(2):87–99, 1989.

[276] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[277] A. Kendon. The consistency of gaze patterns in social interaction. *British Journal of Psychology*, 60(4):481–494, 1969.

[278] A. Kendon. *Sudies in Dyadic Communication*, chapter Some Relationships Between Body Motion and Speech: An Analysis of an Example, pages 177–208. Pergamon Press, 1972.

[279] A. Kendon. Some uses of the head shake. *Gesture*, 2(2):147–182, 2002.

[280] E. Keogh. Exact Indexing of Dynamic Time Warping. In *International Conference on Very Large Data Bases*, pages 406–417, 2002.

[281] E. Keogh and S. Kasetty. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[282] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.

[283] S.-H. Kim, N.-K. Kim, S. C. Ahn, and H.-G. Kim. Object oriented face detection using range and color information. In *International Conference on Automatic Face and Gesture Recognition*, pages 76–81, 1998.

[284] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. Contemporary Mathematics. American Mathematical Society, 1980.

[285] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[286] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[287] R. E. Kleck, R. C. Vaughan, J. Cartwright-Smith, K. B. Vaughan, C. Z. Colby, and J. T. Lanzetta. Effects of Being Observed on Expressive, Subjective, and Physiological

Responses to Painful Stimuli. *Journal of Personality and Social Psychology*, 34(6):1211–1218, 1976.

[288] C. L. Kleinke. Gaze and Eye Contact: A Research Review. *Psychological Bulletin*, 100(1):78–100, 1986.

[289] W. B. Knox and P. Stone. Tamer: Training an agent manually via evaluative reinforcement. In *International Conference on Development and Learning*, pages 292–297, 2008.

[290] H. Kobayashi and F. Hara. Facial interaction between animated 3D face robot and human beings. In *International Conference on Systems, Man, and Cybernetics*, volume 4, pages 3732–3737, 1997.

[291] H. Kobayashi and S. Kohshima. Unique morphology of the human eye. *Nature*, 387:767–768, 1997.

[292] C. Kotropoulos and I. Pitas. Rule-based face detection in frontal views. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2537–2540, 1997.

[293] E. Krahmer, M. Swerts, M. Theune, and M. Weegelsa. The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication*, 36(1–2):133–145, 2002.

[294] R. E. Kraut. Social Presence, Facial Feedback, and Emotion. *Journal of Personality and Social Psychology*, 42(5):853–863, 1982.

[295] R. E. Kraut and R. E. Johnston. Social and Emotional Messages of Smiling: An Ethological Approach. *Journal of Personality and Social Psychology*, 37(9):1539–1553, 1979.

[296] N. Krüger, M. Pötzsch, and C. von der Malsburg. Determination of face position and pose with a learned representation based on labelled graphs. *Image and Vision Computing*, 15(8):665–673, 1997.

[297] V. Krüger and G. Sommer. Gabor Wavelet Networks for Efficient Head Pose Estimation. *Image and Vision Computing*, 20(9–10):665–672, 2002.

[298] H. Kruppa, M. Castrillón Santana, and B. Schiele. Fast and Robust Face Finding via Local Context. In *Joint International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 157–164, 2003.

[299] T. Kudo, E. Maeda, and Y. Matsumoto. An Application of Boosting to Graph Classification. In *Advances in Neural Information Processing Systems (NIPS)*, volume 17, pages 729–736, 2004.

[300] Y. H. Kwon and N. da Vitoria Lobo. Face Detection Using Templates. In *International Conference on Pattern Recognition*, volume 1, pages 764–767, 1994.

[301] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *Transactions on Computers*, 42(3):300–311, 1993.

[302] K.-M. Lam and H. Yan. Fast algorithm for locating head boundaries. *Journal of Electronic Imaging*, 3(4):351–359, 1994.

[303] K.-M. Lam and H. Yan. An Analytic-to-Holistic Approach for Face Recognition Based on a Single Frontal View. *Transactions on Pattern Analysis and Machine Intelligence*, 20(7):673–686, 1998.

[304] C. Lang. Personenidentifikation mit Active Appearance Models. Master's thesis, Bielefeld University, Faculty of Technology, Applied Informatics, 2007.

[305] C. Lang, M. Hanheide, M. Lohse, H. Wersing, and G. Sagerer. Feedback Interpretation based on Facial Expressions in Human-Robot Interaction. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 189–194, 2009.

[306] C. Lang, S. Wachsmuth, M. Hanheide, and H. Wersing. Facial Communicative Signal Interpretation in Human-Robot Interaction by Discriminative Video Subsequence Selection. Technical report, Bielefeld University, Faculty of Technology, Research Institute for Cognition and Robotics, Applied Informatics, 2012.

[307] C. Lang, S. Wachsmuth, M. Hanheide, and H. Wersing. Facial Communicative Signals - Valence Recognition in Task-Oriented Human-Robot Interaction. *International Journal of Social Robotics - Special Issue on Measuring Human-Robot Interaction*, to appear, 2012.

[308] C. Lang, S. Wachsmuth, H. Wersing, and M. Hanheide. Facial Expressions as Feedback Cue in Human-Robot Interaction - a Comparison between Human and Automatic Recognition Performances. In *Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, pages 79–85, 2010.

[309] S. R. H. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Quarterly Journal of Experimental Psychology Section A*, 53(3):825–845, 2000.

[310] S. R. H. Langton and V. Bruce. Reflexive Visual Orienting in Response to the Social Attention of Others. *Visual Cognition*, 6(5):541–567, 1999.

[311] S. R. H. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & Psychophysics*, 66(5):752–771, 2004.

[312] S. R. H. Langton, R. J. Watt, and V. Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59, 2000.

[313] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401, 1995.

[314] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic Interpretation and Coding of Face Images Using Flexible Models. *Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[315] Y. LeCun and F. J. Huang. Loss Functions for Discriminative Training of Energy-Based Models. In *International Workshop on Artificial Intelligence and Statistics*, 2005.

[316] H. Lei and B. Sun. A Study on the Dynamic Time Warping in Kernel Machines. In *International Conference on Signal-Image Technologies and Internet-Based System*, pages 839–845, 2007.

[317] T. K. Leung, M. C. Burl, and P. Perona. Finding Faces in Cluttered Scenes using Random Labeled Graph Matching. In *International Conference on Computer Vision*, pages 637–644, 1995.

[318] M. S. Lew. Information Theoretic View-Based and Modular Face Detection. In *International Conference on Automatic Face and Gesture Recognition*, pages 198–203, 1996.

[319] S. Z. Li, Q. Fu, L. Gu, B. Schölkopf, Y. Cheng, and H. Zhang. Kernel Machine Based Learning For Multi-View Face Detection and Pose Estimation. In *International Conference on Computer Vision*, volume 2, pages 674–679, 2001.

[320] S. Z. Li and Z. Zhang. FloatBoost Learning and Statistical Face Detection. *Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1112–1123, 2004.

[321] S. Z. Li, L. Zhu, Z. Q. Zhang, A. Blake, H. J. Zhang, and H. Shum. Statistical Learning of Multi-view Face Detection. In *European Conference on Computer Vision*, volume 2353/2006 of *Lecture Notes in Computer Science*, pages 117–121, 2002.

[322] Y. Li, S. Gong, J. Sherrah, and H. Liddel. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, 22(5):413–427, 2004.

[323] Z. Li, Y. Fu, J. Yuan, T. S. Huang, and Y. Wu. Query Driven Localized Linear Discriminant Models for Head Pose Estimation. In *International Conference on Multimedia and Expo*, pages 1810–1813, 2007.

[324] W. L. Libby. Eye contact and direction of looking as stable individual differences. *Journal of Experimental Research in Personality*, 4(4):303–312, 1970.

[325] M. D. Lieberman, N. I. Eisenberger, M. J. Crockett, S. M. Tom, J. H. Pfeifer, and B. M. Way. Putting Feelings Into Words: Affect Labeling Disrupts Amygdala Activity in Response to Affective Stimuli. *Psychological Science*, 18(5):421–428, 2007.

[326] J.-J. J. Lien. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity.* PhD thesis, Carnegie Mellon University, Robotics Institute, 1998.

[327] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. Technical report, Microprocessor Research Lab, Intel Labs, 2002.

[328] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *International Conference on Image Processing*, volume 1, pages I–900–I–903, 2002.

[329] S.-H. Lin, S.-Y. Kung, and L.-J. Lin. Face Recognition/Detection by Probabilistic Decision-Based Neural Network. *Transactions on Neural Networks*, 8(1):114–132, 1997.

[330] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The Computer Expression Recognition Toolbox (CERT). In *International Conference on Automatic Face and Gesture Recognition*, pages 298–305, 2011.

[331] Q. Liu and G.-z. Peng. A Robust Skin Color Based Face Detection Algorithm. In *International Asia Conference on Informatics in Control, Automation and Robotics*, pages 525–528, 2010.

[332] K. S. Lohan, A.-L. Vollmer, J. Fritsch, K. Rohlfing, and B. Wrede. Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations? In *International Workshop on Social Signal Processing*, 2009.

[333] M. Lohse. *Investigating the influence of situations and expectations on user behavior: empirical analyses in human-robot interaction.* PhD thesis, Research Institute for Cognition and Robotics, Faculty of Technology, Bielefeld University, 2010.

[334] M. Lohse and M. Hanheide. Evaluating a social home tour robot applying heuristics. In *Workshop Robots as Social Actors at RO-MAN*, 2008.

[335] M. Lohse, K. J. Rohlfing, B. Wrede, and G. Sagerer. Try something else! When users change their discursive behavior in human-robot interaction. In *International Conference on Robotics and Automation*, Pasadena, CA, USA, May 2008.

[336] F. Lömker, S. Wrede, M. Hanheide, and J. Fritsch. Building Modular Vision Systems with a Graphical Plugin Environment. In *Proceedings of the International Conference on Computer Vision Systems*, St. Johns University, Manhattan, New York City, USA, 2006.

[337] F. Lömker *et al.* iceWing - A graphical plugin shell. `http://icewing.sourceforge.net/`, 2005–2012. Visted on 2012-02-28.

[338] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[339] S. Lucey, A. B. Ashraf, and J. F. Cohn. *Face Recognition*, chapter Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face, pages 275–286. I-TECH Education and Publishing, 2007.

[340] J. Lyons. *Introduction to Theoretical Linguistics*. Cambridge University Press, 1968.

[341] Y. Ma, Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade. Sparse Bayesian Regression for Head Pose Estimation. In *International Conference on Pattern Recognition*, pages 507–510, 2006.

[342] M. Malciu and F. Preteux. A robust model-based approach for 3D head tracking in video sequences. In *International Conference on Automatic Face and Gesture Recognition*, pages 169–174, 2000.

[343] D. Matsumoto and C. Kupperbusch. Idiocentric and allocentric differences in emotional expression, experience, and the coherence between expression and experience. *Asian Journal of Social Psychology*, 4:113–131, 2001.

[344] I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision*, 60:135–164, 2004.

[345] D. Maurer. *Social Perception in Infants*, chapter InfantsâĂŹ perception of facedness, pages 73–100. Ablex Publishing, 1985.

[346] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. ELAN - EUDICO Linguistic Annotator. `http://www.lat-mpi.eu/tools/elan/`, 2002–2012. Visted on 2012-02-28.

[347] S. K. Maynard. Interactional functions of a nonverbal sign Head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, 11(5):589–606, 1987.

[348] J. Maynard-Smith and D. Harper. *Animal Signals*. Oxford University Press, 2004.

[349] F. T. McAndrew. A Cross-Cultural Study of Recognition Thresholds for Facial Expressions of Emotion. *Journal of Cross-Cultural Psychology*, 17:211–224, 1986.

[350] G. McCarthy, A. Puce, J. C. Gore, and T. Allison. Face-Specific Processing in the Human Fusiform Gyrus. *Journal of Cognitive Neuroscience*, 9(5):605–610, 1997.

[351] E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7):855–878, 2000.

[352] D. McDuff, R. Kaliouby, K. Kassam, and R. Picard. Affect Valence Inference From Facial Action Unit Spectrograms. In *Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 17–24, 2010.

[353] S. J. McKenna and S. Gong. Real-time face pose estimation. *Real-Time Imaging: Special issue on real-time visual monitoring and inspection*, 4(5):333–347, 1998.

[354] S. J. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17(3–4):225–231, 1999.

[355] E. McKone and R. Robbins. The evidence rejects the expertise hypothesis: Reply to Gauthier & Bukach. *Cognition*, 103(2):331–336, 2007.

[356] J. F. C. McLachlan. A short adjective check list for the evaluation of anxiety and depression. *Journal of Clinical Psychology*, 32(1):195–197, 1976.

[357] D. M. McNaira and M. Lorr. An analysis of mood in neurotics. *Journal of Abnormal and Social Psychology*, 69(6):620–627, 1964.

[358] A. Mehrabian and J. T. Friar. Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology*, 33(3):330–336, 1969.

[359] A. Mehrabian and J. A. Russell. *An Approach to Environmental Psychology*. The MIT Press, Cambridge, MA, US, 1974.

[360] J. Meynet, V. Popovici, and J.-P. Thiran. Face detection with boosted Gaussian features. *Pattern Recognition*, 40(8):2283–2291, 2007.

[361] T. Mita, T. Kaneko, and O. Hori. Joint Haar-like Features for Face Detection. In *International Conference on Computer Vision*, number 2, pages 1619–1626, 2005.

[362] A. Modigliani. Embarrassment, facework, and eye contact: Testing a theory of embarrassment. *Journal of Personality and Social Psychology*, 17(1):15–24, 1971.

[363] C. H. Morimoto and M. R. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005.

[364] S. Morishita. Computing Optimal Hypotheses Efficiently for Boosting. In S. Arikawa and A. Shinohara, editors, *Progress in Discovery Science*, volume 2281/2002 of *Lecture Notes in Computer Science*, pages 218–222, 2002.

[365] M. T. Motley and C. T. Camden. Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Speech Communication*, 52(1):1–22, 1988.

[366] A. Mueen, E. Keogh, and N. E. Young. Logical-Shapelets: an Expressive Primitive for Time Series Classification. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pages 1154–1162, 2011.

[367] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi. Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation. In *Intelligent Transportation Systems Conference*, pages 709–714, 2007.

[368] E. Murphy-Chutorian and M. M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. *Transaction on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[369] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a Model of Face-to-Face Grounding. In *41st Annual Meeting of the Association for Computational Linguistics*, pages 553–561, 2003.

[370] A. V. Nefian and M. H. I. Hayes. Face detection and recognition using hidden Markov models. In *International Conference on Image Processing*, volume 1, pages 141–145, 1998.

[371] Neurotechnology. VeriLook SDK - Face identification for PC or Web applications. `http://download.neurotechnology.com/VeriLook_SDK_Brochure_2012-06-05.pdf`. visited on 2012-06-12.

[372] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. Real-Time Stereo Tracking for Head Pose and Gaze Estimation. In *International Conference on Automatic Face and Gesture Recognition*, pages 122–128, 2000.

[373] J. Ng and S. Gong. Composite support vector machines for detection of faces across views and pose estimation. *Image and Vision Computing*, 20(5–6):359–368, 2002.

[374] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In *British Machine Vision Conference*, volume III, pages 1249–1258, 2006.

[375] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[376] G. Nielsen. *Studies in Self Confrontation.* Munksgaard, 1964.

[377] T. Niit and J. Valsiner. Recognition of facial expressions: An experimental investigation of Ekman's model. *Ada et Commentationes Universitatis Tarvensis*, 429:85–107, 1977.

[378] A. Nikolaidis and I. Pitas. Facial Feature Extraction and Pose Determination. *Pattern Recognition*, 33(11):1783–1791, 2000.

[379] S. Niyogi and W. T. Freeman. Example-Based Head Tracking. In *International Conference on Automatic Face and Gesture Recognition*, pages 374–378, 1996.

[380] S. Nowozin, G. Bakir, and K. Tsuda. Discriminative Subsequence Mining for Action Classification. In *International Conference on Computer Vision*, pages 1–8, 2007.

[381] S. Ohayon and E. Rivlin. Robust 3D Head Tracking Using Camera Pose Estimation. In *International Conference on Pattern Recognition*, volume 1, pages 1063–1066, 2006.

[382] K. Oka, Y. Sato, Y. Nakanishi, and H. Koike. Head Pose Estimation System Based on Particle Filtering with Adaptive Diffusion Control. In *Conference on Machine Vision Applications*, pages 586–589, 2005.

[383] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic Face Detection and Pose Estimation with Energy-Based Models. *Journal of Machine Learning Research*, 8:1197–1215, 2007.

[384] C. Osgood, G. Suci, and P. Tannenbaum. *The Measurement of Meaning.* University of Illinois Press, Urbana, USA, 1957.

[385] C. E. Osgood. Dimensionality of the Semantic Space for Communication via Facial Expressions. *Scandinavian Journal of Psychology*, 7(1):1–30, 1966.

[386] E. Osuna, R. Freund, and F. Girosit. Training Support Vector Machines: an Application to Face Detection. In *Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[387] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005.

[388] T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using HMM. In *International Conference on Automatic Face and Gesture Recognition*, pages 442–447, 1998.

[389] J. P. Otteson and C. R. Otteson. Effect of Teacher's Gaze on Children's Story Recall. *Perceptual and Motor Skills*, 50(1):35–42, 1980.

[390] C. Padgett and G. Cottrell. Representing Face Image for Emotion Classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 9, pages 894–900. MIT Press, 1997.

[391] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of Facial Expression. *Image and Vision Computing Journal*, 18(11):881–905, 2000.

[392] M. Pantic and L. J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[393] B. Parkinson. Do Facial Movements Express Emotions or Communicate Motives? *Personality and Social Psychology Review*, 9(4):278–311, 2005.

[394] P. Patel, E. Keogh, J. Lin, and S. Lonardi. Mining Motifs in Massive Time Series Databases. In *International Conference on Data Mining*, pages 370–377, 2002.

[395] Pattern Recognition and Image Processing / Image Analysis, Department of Computer Science, Faculty of Engineering, University of Freiburg. LIBSVMTL - a Support Vector Machine Template Library. `http://lmb.informatik.uni-freiburg.de/lmbsoft/libsvmtl.en.html`. visited on 2012-06-18.

[396] M. L. Patterson. Stability of nonverbal immediacy behaviors. *Journal of Experimental Social Psychology*, 9(2):97–109, 1973.

[397] M. L. Patterson, S. Mullens, and J. Romano. Compensatory Reactions to Spatial Intrusion. *Sociometry*, 34(1):114–121, 1971.

[398] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, 2004.

[399] D. Perrett and N. Emery. Understanding the Intentions of Others from Visual Signals - Neurophysiological Evidence. *Cahiers de Psychologie Cognitive*, 13(5):683–694, 1994.

[400] V. Peters. Effizientes Training ansichtsbasierter Gesichtsdetektoren. Master's thesis, Bielefeld University, Faculty of Technology, Applied Informatics, 2006.

[401] T. V. Pham, M. Worring, and A. W. M. Smeulders. Face detection by aggregated Bayesian network classifiers. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 249–262, 2001.

[402] R. Plutchik. *Theories of Emotion (Emotion: Theory, Research, and Experience, 1)*, chapter A general psycho-evolutionary theory of emotion, pages 3–33. Academic Press, 1980.

[403] I. Poggi, F. D'Errico, and L. Vincze. Types of Nods. The polysemy of a social signal. In *7th International Conference on Language Resources and Evaluation*, pages 17–23, 2010.

[404] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715–734, 2005.

[405] M. I. Posner, M. J. Nissen, and W. C. Ogden. *Modes of Perceiving and Processing Information*, chapter Attended and unattended processing modes: the role of set for spatial location, pages 137–157. Lawrence Erlbaum Associates, 1978.

[406] J. A. Prado, C. Simplício, N. F. Lori, and J. Dias. Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction. *International Journal of Social Robotics*, 4(1):29–51, 2012.

[407] R. R. Provine and K. R. Fischer. Laughing, Smiling, and Talking: Relation to Sleeping and Social Context in Humans. *Ethology*, 83(4):295–305, 1989.

[408] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning. Morgan Kaufman Publ Inc, 1992.

[409] A. Rabie, C. Lang, M. Hanheide, M. Castrillón-Santana, and G. Sagerer. Automatic Initialization for Facial Analysis in Interactive Robotics. In *Proceedings of the International Conference on Computer Vision Systems*, pages 517–526, Santorini, Greece, May 2008. Springer.

[410] A. Rabie, B. Wrede, T. Vogt, and M. Hanheide. Evaluation and Discussion of Multimodal Emotion Recognition. In *Second International Conference on Computer and Electrical Engineering*, volume 1, pages 598–602, 2009.

[411] B. Raducanu and F. Dornaika. Appearance-based face recognition using a supervised manifold learning framework. In *Workshop on Applications of Computer Vision*, pages 465–470, 2012.

[412] R. Rae and H. J. Ritter. Recognition of human head orientation based on artificial neural networks. *Transactions on Neural Networks*, 9(2):257–265, 1998.

[413] T. M. Rath and R. Manmatha. Lower-Bounding of Dynamic Time Warping Distances for Multivariate Time Series. Technical Report MM-40, Center for Intelligent Information Retrieval, University of Massachusetts Amherst, 2002.

[414] W. Reich. Toward a Computational Model of 'Context'. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*, pages 48–53, 2011.

[415] R. Reisenzein, S. Bördgen, T. Holtbernd, and D. Matz. Evidence for Strong Dissociation Between Emotion and Facial Displays: The Case of Surprise. *Journal of Personality and Social Psychology*, 91(2):295–315, 2006.

[416] P. Ricciardelli, G. Baylis, and J. Driver. The positive and negative of human expertise in gaze perception. *Cognition*, 77(1):B1–B14, 2000.

[417] D. A. Roark, A. J. OâĂŹToole, and H. Abdi. Human Recognition of Familiar and Unfamiliar People in Naturalistic Video. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 36–43, 2003.

[418] R. Robbins and E. McKone. No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, 103(1):34–79, 2007.

[419] M. Roberts, T. Cootes, and J. Adams. Robust Active Appearance Models with Iteratively Rescaled Kernels. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 302–311, 2007.

[420] A. Robotics. Nao. `http://www.aldebaran-robotics.com/en/`. visited on 2012-06-26.

[421] B. Rossion, C.-C. Kung, and M. J. Tarr. Visual expertise with nonface objects leads to competition with the early perceptual processing of faces in the human occipitotemporal cortex. *Proceedings of the National Academy of Sciences USA*, 101(40):14521–14526, 2004.

[422] S. T. Roweis and K. S. Lawrence. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[423] H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[424] H. A. Rowley, S. Baluja, and T. Kanade. Rotation Invariant Neural Network-Based Face Detection. In *Conference on Computer Vision and Pattern Recognition*, pages 38–44, 1998.

[425] J. A. Russell. Evidence of Convergent Validity on the Dimensions of Affect. *Journal of Cross-Cultural Psychology*, 36(10):1152–1168, 1978.

[426] J. A. Russell. Affective Space is Bipolar. *Journal of Personality and Social Psychology*, 37(3):345–356, 1979.

[427] J. A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[428] J. A. Russell. Forced-Choice Response Format in the Study of Facial Expression. *Motivation and Emotion*, 17(1):41–51, 1993.

[429] J. A. Russell. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, 1994.

[430] J. A. Russell. Facial expressions of emotion: what lies beyond minimal universality? *Psychological Bulletin*, 118(3):379–91, 1995.

[431] J. A. Russell and J. M. Fernández-Dols. *The Psychology of Facial Expression*, chapter What does a facial expression mean?, pages 3–30. Cambridge University Press, 1997.

[432] J. A. Russell, M. Lewicka, and T. Niit. A Cross-Cultural Study of a Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 57(5):848–856, 1989.

[433] S. Sakoe, H.; Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[434] Y. Sakurai, S. Papadimitriou, and C. Faloutsos. BRAID: Stream Mining through Group Lag Correlations. In *SIGMOD International Conference on Management of Data*, pages 599–610, 2005.

[435] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition*, 25(1):65–77, 1992.

[436] A. Samal and P. A. Iyengar. Human Face Detection Using Silhouettes. *International Journal of Pattern Recognition and Artificial Intelligence*, 9(6):845–867, 1995.

[437] F. Samaria and S. Young. HMM-based architecture for face identification. *Image and Vision Computing*, 12(8):537–543, 1994.

[438] D. Sander, D. Grandjean, S. Kaiser, T. Wehrle, and K. R. Scherer. Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *European Journal of Cognitive Psychology*, 19(3):470–480, 2007.

[439] J. M. Saragih, S. Lucey, and J. F. Cohn. Face Alignment through Subspace Constrained Mean-Shifts. In *International Conference of Computer Vision*, pages 1034–1041, 2009.

[440] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[441] D. Saxe and R. Foulds. Toward Robust Skin Identification in Video Images. In *International Conference on Automatic Face and Gesture Recognition*, pages 379–384, 1996.

[442] R. E. Schapire and Y. Singer. Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*, 37(3):297–336, 1999.

[443] K. R. Scherer, A. Shorr, and T. Johnstone, editors. *Appraisal Processes in Emotion: Theory, Methods, Research.* Oxford University Press, 2001.

[444] H. Schneiderman and T. Kanade. Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 45–51, 1998.

[445] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning).* MIT Press, 2002.

[446] C. Schüldt, I. Laptev, and B. Caputo. Recognizing Human Actions: A Local SVM Approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

[447] R. Schulz and J. Barefoot. Non-verbal Responses and Affiliative Conflict Theory. *British Journal of Social and Clinical Psychology*, 13(3):237–243, 1974.

[448] S. R. Schweinberger, A. M. Burton, and S. W. Kelly. Asymmetric dependencies in perceiving identity and emotion: Experiments with morphed faces. *Attention, Perception, & Psychophysics*, 61(6):1102–1115, 1999.

[449] S. R. Schweinberger and G. R. Soukup. Asymmetric Relationships Among Perceptions of Facial Identity, Emotion, and Facial Speech. *Journal of Experimental Psychology: Human Perception and Performance*, 24(6):1748–1765, 1998.

[450] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion Recognition Based on Joint Visual and Audio Cues. In *18th International Conference on Pattern Recognition*, volume 1, pages 1136–1139, 2006.

[451] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic Facial Expression Analysis. *Image and Vision Computing*, 25(12):1856–1863, December 2007.

[452] E. Seemann, K. Nickel, and R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *International Conference on Automatic Face and Gesture Recognition*, pages 626–631, 2004.

[453] J. Sergent, S. Ohta, and B. MacDonald. Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, 115(1):15–36, 1992.

[454] S. Shacham, R. Dar, and C. S. Cleeland. The Relationship of Mood State to the Severity of Clinical Pain. *Pain*, 18(2):187–197, 1984.

[455] L. G. Shapiro and G. C. Stockman. *Computer Vision*. Prentice Hall, 2001.

[456] J. Sherrah, S. Gong, and E. J. Ong. Face distributions in similarity space under varying head pose. *Image and Vision Computing*, 19(12):807–819, 2001.

[457] H. Shimodaira, K.-i. Noma, M. Nakai, and S. Sagayama. Dynamic Time-Alignment Kernel in Support Vector Machine. In *European Conference on Speech Communication and Technology*, pages 1841–1844, 2001.

[458] F. Siepmann *et al.* BonSAI - Bielefeld Sensor and Actuator Interface. `http://opensource.cit-ec.de/projects/bonsai`, 2008–2012. Visted on 2012-01-11.

[459] D. Simon, K. Craig D., F. Gosselin, P. Belin, and P. Rainville. Recognition and discrimination of prototypical dynamic expressions of pain and emotions. *Pain*, 135(1–2):55–64, 2008.

[460] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[461] P. Sinha, Y. Ostrovsky, and R. Russel. *Encyclopedia of Perception*, chapter Face Perception, pages 445–449. SAGE Publications, Inc, 2010.

[462] P. Sinha and T. Poggio. I think I know that face ... *Nature*, 384:404, 1996.

[463] S. A. Sirohey. Human Face Segmentation and Identification. Technical Report CS-TR-3176, University of Maryland, Center for Automation Research, Computer Vision Laboratory, 1993.

[464] C. A. Smith and H. S. Scot. *The Psychology of Facial Expression*, chapter A Componential Approach to the meaning of facial expressions, pages 229–254. Cambridge University Press, 1997.

[465] J. G. Snider and C. E. Osgood, editors. *Semantic differential technique*. Aldine, Chicago, 1969.

[466] K. Sobottka and I. Pitas. Face localization and feature extraction based on shape and color information. In *International Conference on Image Processing*, pages 483–486, 1996.

[467] K. Sobottka and I. Pitas. A novel method for automatic face segmentation, facial feature extraction and tracking. *Signal Processing: Image Communication*, 12(3):263–281, 1998.

[468] F. A. Soto and E. A. Wasserman. Asymmetrical interactions in the perception of face identity and emotional expression are not unique to the primate visual system. *Journal of Vision*, 11(3):24, 1–18, 2011.

[469] S. Srinivasan and K. L. Boyer. Head pose estimation using view based eigenspaces. In *International Conference on Pattern Recognition*, volume 4, pages 302–305, 2002.

[470] M. Störring. *Computer Vision and Human Skin Colour*. PhD thesis, Aalborg University, Faculty of Engineering and Science, 2004.

[471] M. Störring, H. J. Andersen, and E. Granum. Physics-based modelling of human skin colour under mixed illuminants. *Robotics and Autonomous Systems*, 35(3–4):131–142, 2001.

[472] C. Storti. *Speaking of India: Bridging the Communication Gap When Working With Indians*, chapter Yes, No, and other Problems, pages 35–76. Nicholas Brealey Publishing, 2007.

[473] K.-K. Sung and T. Poggio. Example-Based Learning for View-Based Human Face Detection. *Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[474] Z. S. Tabatabaie, R. W. Rahmat, N. I. B. Udzir, and E. Kheirkhah. A Hybrid Face Detection System using combination of Appearance-based and Feature-based methods. *International Journal of Computer Science and Network Security*, 9(5):181–185, 2009.

[475] B. Takács. Face Location Using a Dynamic Model of Retinal Feature Extraction. In *International Workshop on Automatic Face and Gesture Recognition*, pages 243–247, 1995.

[476] J. W. Tanaka and M. J. Farah. Parts and Wholes in Face Recognition. *The Quarterly Journal of Experimental Psychology*, 46A(2):225–245, 1993.

[477] J. W. Tanaka and J. A. Sengco. Features and their configuration in face recognition. *Memory & Cognition*, 25(5):583–592, 1997.

[478] K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66(1):170–189, 1991.

[479] H. Tao and T. S. Huang. Connected vibrations: a modal analysis approach for non–rigid motion tracking. In *Conference on Computer Vision and Pattern Recognition*, pages 735–740, June 1998.

[480] M. J. Tarr. *Perception of Faces, Objects and Scenes. Analytic and Holistic Processes*, chapter Visual Object Recognition: Can A Single Mechanism Suffice?, pages 177–211. Oxford University Press, 2003.

[481] M. ter Maat and D. Heylen. *Multimodal Signals: Cognitive and Algorithmic Issues*, chapter Using Context to Disambiguate Communicative Signals, pages 67–74. Springer Berlin Heidelberg, 2009.

[482] J.-C. Terrillon, M. David, and S. Akamatsu. Detection of human faces in complex scene images by use of a skin color model and of invariant Fourier-Mellin moments . In *International Conference on Pattern Recognition*, volume 2, pages 1350–1355, 1998.

[483] D. Terzopoulos and K. Waters. Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. *Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[484] R. E. Thayer. Measurement of activation through self-report. *Psychological Reports*, 20(2):663–678, 1967.

[485] Y.-L. Tian, L. Brown, C. Connell, S. Pankanti, A. Hampapur, A. Senior, and R. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 92–99, 2003.

[486] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing Action Units for Facial Expression Analysis. *Transaction on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[487] M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[488] A. Tiwari, R. Gupta, and D. Agrawal. A Survey on Frequent Pattern Mining: Current Status and Challenging Issues. *Information Technology Journal*, 9(7):1278–1293, 2010.

[489] J. L. Tracy and D. Matsumoto. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences of the United States of America*, 105(33):11655–11660, 2008.

[490] D. Y. Tsao, W. A. Freiwald, R. B. H. Tootell, and M. S. Livingstone. A Cortical Region Consisting Entirely of Face-Selective Cells. *Science*, 311(5761):670–674, 2006.

[491] D. Y. Tsao and M. S. Livingstone. Mechanisms of Face Perception. *Annual Review of Neuroscience*, 31:411–437, 2008.

[492] D. Tsishkou, M. Hammami, and L. Chen. Face detection in video using combined data-mining and histogram based skin-color model. In *International Symposium on Image and Signal Processing and Analysis*, volume 1, pages 500–503, 2003.

[493] A. Tsukamoto, C.-W. Lee, and S. Tsuji. Detection and pose estimation of human face with synthesized image models. In *International Conference on Pattern Recognition*, volume 1, pages 754–757, 1994.

[494] K. Tsunoda, Y. Yamane, M. Nishizaki, and M. Tanifuji. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4(8):832–838, 2001.

[495] C. Turati. Why Faces Are Not Special to Newborns: An Alternative Account of the Face Preference. *Current Directions in Psychological Science*, 13(1):5–8, 2004.

[496] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[497] P. E. Valk, D. L. Bailey, D. W. Townsend, and M. N. Maisey, editors. *Positron Emission Tomography: Basic Science and Clinical Practice*. Springer, 2005.

[498] M. F. Valstar, H. Gunes, and M. Pantic. How to Distinguish Posed from Spontaneous Smiles using Geometric Features. In *International Conference on Multimodal Interfaces*, pages 38–45, 2007.

[499] T. A. van Dijk. Discourse, context and cognition. *Discourse Studies*, 8(1):159–177, 2006.

[500] A. C. Varchmin, R. Rae, and H. Ritter. Image based recognition of gaze direction using adaptive methods. In *Gesture and sign language in human-computer interaction. International Gesture Workshop*, Bielefeld, Germany, September 1997.

[501] S. P. Vecera and M. H. Johnson. Gaze detection and the cortical processing of faces: Evidence from infants and adults. *Visual Cognition*, 2(1):59–87, 1995.

[502] E. Vidal, F. Casacuberta, J. M. Benedi, M. J. Lloret, and H. Rulot. On the verification of triangle inequality by dynamic time-warping dissimilarity measures. *Speech Communication*, 7(1):67–79, 1988.

[503] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Conference on Computer Vision and Pattern Recognition*, pages I–511–I–518, 2001.

[504] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[505] M. Voit, K. Nickel, and R. Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. In *International Evaluation Conference on Classification of Events, Activities and Relationships*, pages 291–298, 2006.

[506] A.-L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, and J. Fritsch. People Modify Their Tutoring Behavior in Robot-Directed Interaction for Action Learning. In *International Conference on Development and Learning*, pages 1–6, 2009.

[507] M. von Grünau and C. Anston. The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, 24(11):1297–1313, 1995.

[508] H. L. Wagner, C. J. MacDonald, and A. S. R. Manstead. Communication of Individual Emotions by Spontaneous Facial Expressions. *Journal of Personality and Social Psychology*, 50(4):737–743, 1986.

[509] K. Walker, T. Cootes, and C. Taylor. Automatically building appearance models from image sequences using salient features. *Image and Vision Computing*, 20(5–6):435–440, 2002.

[510] J.-G. Wang and E. Sung. Study on Eye Gaze Estimation. *Transactions on Systems, Man, and Cybernetics*, 32(3):332–350, 2002.

[511] J.-G. Wang and E. Sung. EM enhancement of 3D head pose estimated by point at infinity. *Image and Vision Computing*, 25:1864–1874, 2007.

[512] S. Wang and A. Abdel-Dayem. Improved Viola-Jones Face Detector. In *Taibah University International Conference on Computing and Information Technology*, pages 123–128, 2012.

[513] Y. Wang, S. Lucey, and J. Cohn. Enforcing Convexity for Improved Alignment with Constrained Local Models. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 1998.

[514] O. M. Watson. *Proxemic Behavior: a Cross-Cultural Study.* Mouton De Gruyter, 1970.

[515] P. Watzlawick, J. B. Bavelas, and D. D. Jackson. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes.* W. W. Norton & Company, first edition, 1967.

[516] J. Weizenbaum. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9(1):36–45, 1966.

[517] M. T. Westbrook. Positive affect: A method of content analysis for verbal samples. *Journal of Consulting and Clinical Psychology*, 44(5):715–719, 1976.

[518] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward Practical Smile Detection. *Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009.

[519] S. C. Widen and J. A. Russell. Descriptive and Prescriptive Definitions of Emotion. *Emotion Review*, 2(4):377–378, 2010.

[520] Willow Garage, Intel, and others. OpenCV (Open Source Computer Vision). `http://opencv.willowgarage.com/wiki/`. visited on 2012-06-18.

[521] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: a Professional Framework for Multimodality Research. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559, 2006.

[522] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-Time Tracking of the Human Body. *Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[523] B. Wu, H. Ai, C. Huang, and S. Lao. Fast Rotation Invariant Multi-View Face Detection Based on Real Adaboost. In *International conference on Automatic Face and Gesture Recognition*, pages 79–84, 2004.

[524] J. Wu and M. M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138–1158, 2008.

[525] G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley Series in Pure and Applied Optics. John Wiley & Sons, 2nd edition, 2000.

[526] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana. Fast Time Series Classification using Numerosity Reduction. In *Twenty-Third International Conference on Machine Learning*, pages 1033–1040, 2006.

[527] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-Time Combined 2D+3D Active Appearance Models. In *Conference on Computer Vision and Pattern Recognition*, pages 535–542, 2004.

[528] R. Xiao, L. Zhu, and H.-J. Zhang. Boosting Chain Learning for Object Detection. In *International Conference on Computer Vision*, volume 1, pages 709–715, 2003.

[529] Y. Xu. Revisiting the Role of the Fusiform Face Area in Visual Expertise. *Cerebral Cortex*, 15(8):1234–1242, 2005.

[530] Y. Xu, J. Liu, and N. Kanwisher. The M170 is selective for faces, not for expertise. *Neuropsychologia*, 43(4):588–597, 2005.

[531] G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[532] J. Yang and A. Waibel. A Real-Time Face Tracker. In *Workshop on Applications of Computer Vision*, pages 142–147, 1996.

[533] M.-H. Yang and N. Ahuja. Detecting Human Faces in Color Images. In *International Conference on Image Processing*, volume 1, pages 127–130, 1998.

[534] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *Transaction on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[535] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas. Facial Expression Recognition Based on Dynamic Binary Patterns. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[536] R. Yang and Z. Zhang. Model-Based Head Pose Tracking With Stereovision. In *International Conference on Automatic Face and Gesture Recognition*, 2002.

[537] W. Yang, C. Sun, and L. Zhang. Face Recognition Using a Multi-manifold Discriminant Analysis Method. In *International Conference on Pattern Recognition*, pages 527–530, 2010.

[538] L. Ye and E. Keogh. Time Series Shapelets: A New Primitive for Data Mining. In *International Conference on Knowledge Discovery and Data Minining (ACM SIGKDD)*, pages 947–956, 2009.

[539] L. Ye and E. Keogh. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1–2):149–182, 2011.

[540] M. Yeasin, B. Bullot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *Transactions on Multimedia*, 8(3):500–508, 2006.

[541] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3D Facial Expression Database For Facial Behavior Research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

[542] R. K. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1):141–145, 1969.

[543] M. Yoneyama, A. Ohtake, Y. Iwano, and K. Shirai. Facial expressions recognition using discrete Hopfield neuralnetworks. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 117–120, 1997.

[544] D. H. Yoo, J. H. Kim, B. R. Lee, and M. J. Chung. Non-contact Eye Gaze Tracking System by Mapping of Corneal Reflections. In *International Conference on Automatic Face and Gesture Recognition*, pages 94–99, 2002.

[545] A. W. Young, D. Hellawell, and D. C. Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987.

[546] K. C. Yow and R. Cipolla. A probabilistic framework for perceptual grouping of features for human face detection. In *International Conference on Automatic Face and Gesture Recognition*, pages 16–21, 1996.

[547] A. L. Yuille, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *Conference on Computer Vision and Pattern Recognition*, pages 104–109, 1989.

[548] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[549] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. Huang. Audio-visual affect recognition in activation-evaluation space. In *International Conference on Multimedia and Expo*, 2005.

[550] C. Zhang, R. Hamid, and Z. Zhang. Taylor Expansion Based Classifier Adaptation: Application to Person Detection. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[551] C. Zhang and Z. Zhang. A Survey of Recent Advances in Face Detection. Technical Report MSR-TR-2010-66, Microsoft Research, 2010.

[552] D. Zhang, W. Zuo, D. Zhang, and H. Zhang. Time Series Classification Using Support Vector Machine with Gaussian Elastic Metric Kernel. In *International Conference on Pattern Recognition*, pages 29–32, 2010.

[553] H. Zhang and D. Zhao. Spatial Histogram Features for Face Detection in Color Images. In *Conference on Advances in Multimedia Information Processing*, volume I, pages 377–384, 2004.

[554] Z. Zhang, Y. Hu, M. Liu, and T. Huang. Head pose estimation in seminar room using multi view face detectors. In *International Evaluation Conference on Classification of Events, Activities and Relationships*, pages 299–304, 2006.

[555] G. Zhao, L. Chen, J. Song, and G. Chen. Large Head Movement Tracking Using SIFT-Based Registration. In *15th International Conference on Multimedia*, pages 807–810, 2007.