Universität Bielefeld

Audio-Visual Emotion Recognition For Natural Human-Robot Interaction

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

vorgelegt von

Ahmad Rabie

an der Technischen Fakultät der Universität Bielefeld

15. März 2010

Dipl.-Ing. Ahmad Rabie
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld

Abdruck der genehmigten Dissertation zur Erlangung
des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)
Technische Fakultät der Universität Bielefeld
am 15.03.2010 vorgelegt von Ahmad Rabie
am 31.01.2011 verteidigt und genehmigt

Gutachter:

Prof. Dr.-Ing. Franz Kummert
Prof. Dr.-Ing. Modesto Castrillón-Santana
Dr.-Ing. Marc Hanheide

Prüfungsausschuss:

Prof. Dr.-Ing Franz Kummert
Prof. Dr.-Ing Modesto Castrillón-Santana
Prof. Dr. Philipp Cimiano
Dr. Hendrik Koesling

gedruckt auf alterungsbeständigem Papier nach ISO 9706

Abstract

Since the computer has conquered our life and became an essential need rather than an accessory, more-sophisticated human-computer interaction, beyond the traditional keyboard, mouse, and monitor, is aimed to enable the users to interact with computers more socially. Emotional interaction play a major role in social life, thus the affective human-robot interaction has evolved significantly throughout the last decades. The aim of this thesis is to provide the ability of emotion understanding for a robot. Throughout the thesis, a discrete theory of emotions is used as a frame of reference. According to it, emotions can be classified into some basic emotion classes.

The research is orginized around two goals. The first goal is to enable a robot to infer the emotional state of its interaction partner by analysing the displayed facial expression in non-constrained conditions. To achieve that, a robust, fully automatic, non-invasive, and real-time applicable vision-based system is developed with the ability to be implemented in the robot.

As the aim is to enable the robot to interact with its interactant in real-world scenarios, sitautions in which the user is engaged in conversational sessions present farther challange for such systems. The second goal of this work is to combine facial expression and speech information cues in such a way, as to enable the affective system of the robot to fit such situations. En route to this goal possible affects of facial configuration related to speech on inferring emotions from facial expression is investigated. The results suggest a degraded performance when facial expressions are displyed during speech as displying them deliberately. In order to smooth this effect, information of audio signal is taken into account. The performance of the emotion recognition system is relatively enhanced by fusing facial expression cues and speech information ones into a bimodal system. The perfomance of the bimodal system still, however, degraded comparing with the perfomance stand-alone facial expression analyis system in the case of displayning facial expression deliberately.

Finally, the extent of recognizing each emotion by utilizng each modality is invistagted. The results indicate a highly varying performance of each modality ith the respective emotion class, and for the bimodal system, each modality should be weighted according to its discriminative power for a specific emotion.

Acknoldgment

Writing a doctoral thesis from start to end is a sizeable task. This task would not have been possible without the help and support of many people. Here is the place to thank them.

First of all, I am greatly indebted to my supervisor and friend, Marc Hanheid. He continuously encouraged and trusted me to develop my own ideas, while ensuring that I was following a reasonable path.

I'm also very grateful to Franz Kummert and Modesto Castrillón-Santana for reviewing my thesis and attending my defense.

Special thanks to Gerhard Sagerer for giving me the oppertunity to write my doctoral thesis in the Applied Informatics Group at Bielefeld University. Thanks to all past and present members of the Applied Informatics Group at Bielefeld University for an always rewarding and pleasant atmosphere. I would thank the help and the patient of AGAI members; Niklas Beuter, Lars Schillingman and Marko Tscherepanow, who supported me and whom I often disturbed. Special thanks to my friend Christian Lang. I very much enjoyed the work with you.

Finally, I would like to thank cordially my parents, my brothers and sisters, as well as my small family: Islam, Fatima and Mustafa for their constant patient, love, support and encouragement. Without them this work would also have been impossible. I cannot begin to thank them enough, and they will always have my respect and love.

Ahmad Rabie
April 2011

Contents

1 Introduction **2**

2 Emotion Theory **6**
2.1 Emotion in Human-Human Interaction 7
2.2 Emotion Categorization 8
2.2.1 Discrete Emotions 9
2.2.2 Dimensional Models of Emotions 10
2.2.3 Other Models of Emotions 11
2.3 Emotion Encoding 12
2.3.1 Facial Expression 13
2.3.2 Speech 15
2.3.3 Facial Expression during Speech 16
2.3.4 Internal Physiological Changes 17
2.4 Accuracy of Decoding Other's Emotion 18
2.5 Summary 19

3 Emotion In Human-Robot Interaction **22**
3.1 Affective Computing in Use 24
3.1.1 Robots with Social Abilities 24
3.1.2 BIRON, Social Interactive Robot 26
3.2 Enabling Affective Interaction with BIRON 28

4 Facial Expression Analysis for HRI **29**
4.1 Related Work 32
4.2 Structure of Face analysis system 32
4.3 Face Detection 33
4.3.1 Basic Approaches to Face Detection 34
4.3.2 Selected Face Detection approach 35
4.4 Facial Feature Extraction 36
4.4.1 Geometric-Based Facial Feature Extraction 37
4.4.2 Appearance-Based Facial Feature Extraction 39
4.4.3 Hybrid Methods of Facial Feature Extraction 39
4.4.4 Active Appearance Models, Feature Extractor 40
4.5 Classification 43
4.5.1 Static Approaches 44
4.5.2 Dynamic Approaches 48
4.6 Contribution 49

4.7 Integration Concept in BIRON . . . 52
4.8 Summary . . . 54

5 Audio-Visual Emotion Recognition 55
5.1 Related Work . . . 58
5.2 Emotion Recognition from Speech . . . 59
5.3 Fusion of Multisensory Data for Emotion Recognition . . . 63
5.4 Contribution . . . 65
5.5 Integration Concept in BIRON . . . 68
5.6 Summary . . . 69

6 Evaluation and Discussion 70
6.1 Emotional Databases . . . 70
6.1.1 Self-Report Approach . . . 71
6.1.2 Judgment Approach . . . 71
6.1.3 Facial Configuration-based Approach . . . 72
6.1.4 Reliability of Labeling method . . . 72
6.2 Databases with Emotional Contents . . . 73
6.2.1 Emotional Databases in Use . . . 73
6.2.2 DaFEx Database . . . 75
6.3 Evaluation of Facial-Expression-Based Emotion Analysis System . . . 76
6.4 Evaluation of Speech-Information-Based Emotion Analysis System . . . 80
6.5 Evaluation of Audio-Visual System . . . 81
6.6 Evaluation in Real-Life Conditions . . . 83
6.7 General Discussion . . . 86

7 Conclusion and Future Work 88

8 Appendix 91
8.1 Evaluation of Visual-Based System Using NN . . . 91
8.2 Home-Tour Scenario . . . 92
8.3 Object-Teaching Scenario . . . 92
8.4 Notations . . . 93

Bibliography 94

List of Figures

1.1 A robot engaged in a multimodal dialogue situation with multiple interaction partners 3

2.1 The structure of Emotion according to the organon model proposed by Bühler. The figure is extracted from the work of Hess [61] 8
2.2 Two well-known theories of basic emotion family. (a) Color-wheel-like location of the eight primary emotions proposed by Plutchik [115], and (b) six prototypes presenting the six basic emotions according to Ekman [42]. From left to right and top to bottom: anger, fear, disgust, surprise, happiness, and sadness 10
2.3 Dimensional models of emotions. (a) two-dimensional valence-arousal judgment space proposed by [124], (b) three-dimensional valence-arousal-dominance judgment space proposed by [15] 11
2.4 Position of each possible affective measurement according to its significance in affective human-human interaction and voluntariness. Derived from Partala [112] 14

3.1 Basic architecture of an Affective Computing framework. The figure depicts the three basic components of such a framework; (I)- Sensing the affective state of the user, (III)- adapting according either to the need of the user or to the cognitive architecture of the system[1] and, (II)- Modeling affective behavior of both the user (user affective profile) and the system (cognitive architecture). Derived from Hudlicka [66] 23
3.2 Examples of robots with social skills in research, from upper left to lower right: Leonardo, Kismet, Flobi, and Barthoc 25
3.3 Physical Characteristics of the robot, BIRON. 27

4.1 Facial expression in interaction between human and robot. Interactant's smile can be understood as an acceptance of what the robot has done and trigger a suitable behavior 30
4.2 Special constrained conditions of facial expression analysis systems. (a) Two cameras are mounted to an arrangement that keeps the head fixed in order to enable the system to capture specific views of the face. (b) The face is labeled with colored markers; they are tracked to infer some emotion-related displacements. 31

4.3 Schematic architecture of the facial-expression-based emotion analysis system. The first stage serves for capturing input images, and finding or estimating the location of the face in these images. The middle stage uses the information provided from the first stage to extract some facial features related to the displaying of emotion. These features are finally labeled with one of a predefined number of basic emotions or action units in the facial expression analysis stage 33
4.4 Examples presenting the performance of the exploited face detector in single-user and multiple-user situations; (a) and (b) respectively.The bounding box of the face and the positions of the eyes, nose, and mouth are colored with either green or blue. Green indicates that the detected object is a frontal face, while blue indicates the using of tracking rather than simple detecting. Printed by courtesy of Castrillón [22] 36
4.5 Training an AAM. (a)- An image example of annotated set for training an AAM, the blue points indicate fiducial landmarks. (b)- Triangulation method used to warp each image in the training set to match the base shape 41
4.6 AAM fitting algorithm can fail if rigid head movement, exp., from the position in (a) to the position in (b), is encountered. The images are extracted from the DaFEx database [11]. 43
4.7 An example of mapping into a high-dimensional space. While the random samples in the two-dimensional space are not linearly separable (a), they can be more easily separated after mapping them in a three-dimensional space (b). 47
4.8 Schematic Architecture of the proposed facial analysis system. Positions of the face and some facial features are extracted in the first stage, left. Coordinations of these features are then used to align an AAM, middle. Parameter vectors, which are extracted by AAM, are then classified by a SVM model, right 50
4.9 Face and facial element detection results for some samples of a sequence extracted from DaFEx [11]. 50
4.10 Initialization based on face bounding box and BFFs (first from left) and landmark matching via AAM search (second) for an image from DaFEx [11]. In cases where the initialization is too poor (third), the AAM search algorithm cannot eventually find a correct matching (fourth). 51
4.11 Integration of facial-expression-based emotion analysis system in BIRON. Components 1-4 present the basic structure of the face memory model [58]. The fifth one presents the desired enhancement of providing an emotional ability. 52
4.12 Two work cases of a facial expression recognition system. FE: Facial Expression, ID Person Identity 53

5.1 The six basic emotions are conveyed by facial configurations in two conditions; deliberative (experience an emotion and displaying it without speaking), experience and displaying during speech session. The first and the third rows present the deliberative ones displayed by several individuals, while second and fourth rows present the same emotion of the same individual during speech. The judgment of the displayed emotion is confused in rows two and four if only the facial expressions are considered. Images are extracted from DaFEx database [11]. 57
5.2 Emotion related variations in both original audio signals (left column), and the corresponding 12 MFCCs (right column). The original signals are presented in the time domain; x-axis presents duration in msec, and y-axis presents the amplitude. Variations related to anger, happiness, neutral, and sadness are presented here from top to bottom. The sentence pronounced in each utterance was in Italian *"In quella piccola stanza vuota c'era però soltanto una sveglia"*; in that little empty room there was only an alarm clock. Dafex database [11] 62
5.3 Schematic Architecture of Acoustic-based Emotion Analysis System . . . 63
5.4 Three basic fusion methods used in the current multimodal emotion recognition systems 65
5.5 An example of simple Bayes net with four random variables. The arcs encode the conditional dependencies between the variables. The example is derived from [128], P.p, 627 66
5.6 The structure of the Baysian network used to fuse cues of both unimodals. Evidence of observable nodes − acoustic and visual − is fed as input into the corresponding node. The posteriori probabilities of the unobservable node are computed, with gives fusion as the final result . . . 67
5.7 Integration concept of both unimodals as a bimodal one in BIRON. Each system provides its own decision, which then fused together in the final decision 68

6.1 Six basic emotions presented by six different individuals; extracted from the DaFEx database [11]. The displayed emotions are, from left to right and top to bottom: angry, disgust, fear, happiness, sadness, and surprise . 76
6.2 Examples of image data captured by robot's camera directly, anger and happiness are displayed in left and right image respectively 84

1 Introduction

Human-human communication and interaction have been around since the beginning of humanity. Social interaction information is exchanged through the medium of verbal signs as speech signals and language and non-verbal signs such as gesture, speech tones, and facial expressions. Emotional interaction should play a major role in human-human natural communication. Since the time of Darwin *"Describing laughter: The sound is produced by a deep inspiration followed by short, interrupted, spasmodic contractions of the chest, and especially the diaphragm... the mouth is open more or less widely, with the corners drawn much backwards, as well as a little upwards; and the upper lip is somewhat raised. "*, a large body of researchers has focused on how people encode their emotion, how they decode other's emotions, and which role emotions play in social human-human interaction [42, 61, 77, 124, 135].

The new scientific understanding of emotions on the one hand, and the rapid evolution of computing system skills on the other, provided inspiration to numerous researchers to build machines that will have the ability to recognize, express, model, and communicate emotions. Rosalind Picard's book triggered an explosion of interest in the emotional side of computers. Consequently, a new research area called *"affective computing"* emerged. Affective computing advocates the idea that emotions are not only useful, but rather required when building truly intelligent computing systems is being aimed at. Thus, Picard suggested that it might be essential for machines to posses either some or all the emotional intelligence and skills humans do [114].

An increasing number of scientists in the field of computer science, inspired by the results of the theoretical studies mentioned above, have focused on mirroring human-human interaction in the field of human-computer interaction (HCI), and more recently human-robot interaction(HRI). Approaches to fully automatic recognition of emotions have emerged, as the necessity for dealing with the affective state of a user has become obvious for efficient and user-friendly human-robot interaction [114]. For example, in tutoring systems or computer games, knowing about the user's feeling of boredom, frustration or happiness can increase learning success or fun in the game [73]. Driving assistant systems will benefit from inferring the pilot/driver's level of confusion in order to avoid possible accidents [59]. In human-robot interaction, affective reactions of the robot, following the recognition of the user's emotional state, can make the interaction more natural and human-like [60, 151].

Toward realizing such interaction, the initial focus was on automatic facial expression analysis, and more precisely, on the recognition of the prototypical emotions from posed static input. Almost all the work from the early 1990s attempted to recognize prototypical emotions from two static face images: neutral and expressive [129]. In the second half of the 1990s, automated face expression analysis started focusing on posed video sequences

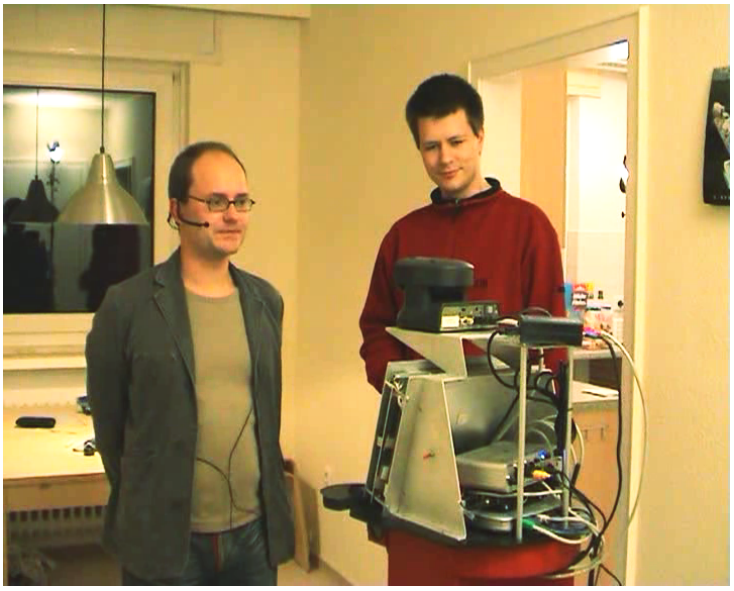


Figure 1.1: *A robot engaged in a multimodal dialogue situation with multiple interaction partners.* Both visual and acoustic information are considered.

and exploiting temporal information in the displayed face expressions [97]. In parallel to the automatic emotion recognition from visual input, works focusing on audio input, physiological measurements, and body language emerged [32, 133].

One major limitation of affective computing is that most of the past research focused on emotion recognition from a single sensorial source, or modality. However, as natural human-human interaction is multimodal, the single sensory observations are often ambiguous, uncertain, incomplete, and influenced by other modalities. A simple solution that overcomes this problem is to utilize a combination of multiple modalities for emotion recognition. This combination in turn triggered other research areas, such which measurements have to be combined and how to combine them.

Fruitful avenues of combining several measurements can be found in the literature. Kim [75] suggested a combination of speech information and physiological measurements. Other scientists preferred to combine speech information with textual content [24], facial expression with physiological measurements [73], while other researchers advocated that a reliable automatic affect recognition system should attempt to combine face expressions and body gestures [70, 74].

For human beings, however, facial expression and voice reveal a person's emotion the most, as will be discussed in Sec. 2.4. Furthermore, a human perceives and understands another's emotion in a multimodal, rather than unimodal way. Indeed the combination of audio and visual information provides more reliable estimates of emotional states. The complementary relationship of these two modalities makes the inference of emotion more accurate than only using a single modality. Acting on this fact in designing emotion analysis systems, a combination of facial expression and speech tone information is the most suitable way towards natural and non-invasive human-robot interaction. Following that, most researchers in this field adopted this kind of combination by designing reliable affective systems [90, 140, 172, 171]. That does not mean, however, that the research in

this field has reached its final aim of having emotion analysis systems that perform equally well as human beings.

Another obstacle challenges both unimodal and multimodal systems when restricting the focus to joint visual and acoustic modalities, namely dealing with real-life scenarios. An example of such scenarios is situations in which the robot and the user are engaged in conversational sessions. The mutual influence between facial configurations that reflect the internal affective states and those caused by speech production processes (movements in the upper and lower part of the face) is seldom challenged.

Yet another challenge is the system's real-time and fully automatic applicability. Indeed applying such systems in real-world scenarios demands that these systems perform well in real-time and fully automatic. That if the former is lacking could delay the reaction of the robot and consequently lead the interaction to be cold, incompetent, and socially inept, while neglecting the latter causes the interaction to be something far removed from the natural one intended.

When employing an emotion analysis system in the scenarios of interaction with multi-person, person-dependent and person-independent systems present another challenge that should be considered. Most existing literature on automatic emotion recognition has not dealt this point. Anecdotal evidence suggests that a person-dependent system outperforms a person-independent one. That is because, considering the human face, the latter describes the geometrical variations in the shape of the face, rather than describing the configurations within the encountered face; these changes are better described by the former system.

The main goal of this thesis can be summarized into the following research questions that will be answered in the remainder of the work.

- When aiming at natural human-robot interaction which cues should be considered and why?
  Facial-expression and speech information are used in human-human interaction and should be considered due to their naturality, low level of voluntariness, and being non-invasive in contrast to the internal physiological measurements

- How does the system behave when the robot and its interaction partner are engaged in a conversational session?
  The performance of a facial expression analysis system is expected to be degraded because of the difficulty of distinguishing between facial configuration related to emotion and that related to speech production processes

- When multimodal human-robot interaction is considered, how should these modalities be fused, and why?
  To smooth the effect of speech-related configurations of the face on inferring emotions from facial expression, speech information can be considered in a complementary rather than conflicting way. The emotion inferred from facial expressions and speech information can be fused, so that the overall performance of the system is enhanced compared to analyzing both stand-alone modalities

- Is there a relation between class-dependent recognition performance and choice of modality? The results indicate that the performance of each modality is highly varying with the respective emotion class.

Our work in this thesis aims at contributing to the development toward an ideal emotion analysis system that enables a robot to behave well in emotional real-life human-robot interaction (HRI). The next chapter starts by discussing several definitions of emotions, three types of theories on emotions, how the emotions can be encoded and decoded, through which modality emotions are presented best, and the multimodality of emotion experience and perception.

In chapter 3 the term affective computing will be introduced, and some application fields will be discussed as well. The chapter will introduce our mobile robot as well as its behavior in social human-robot interaction situations.

A real-time fully automatic facial-expression-based emotion analysis system will be introduced in chapter. 4. In contrast to most facial expression analysis systems which are currently used, the proposed system in this chapter fulfils most of the requirements of an ideal system; these requirements will be discussed intensively in the same chapter.

In order to enhance the performance of the system in human-robot conversational situations, facial expressions and vocal information are fused, yielding a fully automatic real life bimodal emotion analysis system. The proposed bimodal system and the fusion method used to build it will be discussed in chapter 5.

A comprehensive evaluations of each stand-alone uni-modals as well the bimodal system on a convenient database are included in Chapter. 6. As the focus of our work is to give these systems an online ability to be employed in life-like human-robot interaction an evaluation is also conducted on data captured in real-life conditions, more information can be found in Chapter. 6 too. A Conclusion and outlook will conclude our thesis.

2 Emotion Theory

> *" An emotion is a conscious mental reaction (such as anger or fear) subjectively experienced as a strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body."*
>
> Merriam Webster Online Dictionary[1]

The concept of what emotions are is very hard to comprehend because emotions are not clearly defined. A psychologist would definitely give a different definition to that of a linguist, a computer scientist or an average person. Emotions as a concept have a long and stressful history. They have been studied since antiquity by philosophers and psychologists. Darwin demonstrated that some emotions, referred to as primary emotions, represent universal emotional processes useful for survival [35]. This aspect was considered as the first of the four emotion perspectives demonstrated in [120]. The second is called the Jamesian perspective, which adopted that the emotional experience is largely due to the experience of bodily changes. The cognitive perspective underlines that the cognitive appraisals of the environment are the underlying causal explanations for emotional processes. Finally, the social perspective emphasizes the importance of culture and context in understanding what occurs in society [120].

Kleinginna and Kleinginna had already recorded around one hundred definitions presented in scientific literature. Most of them behold emotions from only one aspect or only one subset of what is generally considered as emotion [77]. The following definitions, which were adopted by Oatley et.al. [101], have often been quoted and are considered as being accepted by the researchers of this field.

(i) An emotion is usually caused by a person consciously or unconsciously evaluating an event as relevant to a concern (a goal) that is important; the emotion is felt as positive when a concern is advanced and negative when a concern is impeded.

(ii) The core of an emotion is readiness to act and the prompting of plans; an emotion gives priority to one or a few actions to which it gives a sense of urgency so it can interrupt, or compete with, alternative mental processes or actions. Different types of readiness create different outline relationships with others.

(iii) An emotion is usually experienced as a distinctive type of mental state, sometimes accompanied or followed by bodily changes, expressions and actions.

Mirroring these definitions in the field of human-robot interaction, the event that has to be evaluated by the robot is some verbal or non-verbal cues from the interactant which are

[1] http://www.merriam-webster.com/dictionary

associated the experience of emotion, where the adaption of the robot according to these changes can be considered as the desired reaction of the robot.

Other emotion-related concepts have sparked the interest of researchers into describing the psychological and physiological phenomena accompanying them. While it is agreed that emotions are considered as being short term, consciously perceived, a valanced state; either positive or negative, like e.g. happy, angry or sad, a mood can last one or several days, weeks, or even months, such as when one is cheerful or depressed. Being often elicited by an internal, or external emotion trigger, targeting either cognitive or social behavior the emotions differentiate from the mood which lack a specific target as well a specific trigger. What also has to be recognized is the difference between emotion and feeling such as (liking or hating), which is referred by Damasio as a private, mental experience of an emotion. According to him, feelings do not include bodily emotional responses, but merely mental perceptions of the state of the body [34].

What follows in this chapter will provide summarized answers to some key questions related to emotion, emotion expression and emotional human-human interaction. These answers will serve as the background to our work presented in this thesis. Section 2.1 will provide an overview of the role of emotion in human-human interaction. Several models of emotion categorization will be discussed in section 2.2. Which cues could be conveyed when we experience emotion, how and how accurately do humans perceive another's emotion will be discussed in sections 2.3 and 2.4 respectively. A small summary will conclude this chapter.

2.1 Emotion in Human-Human Interaction

In order to sustain a social relationship the most effective human-human interactive medium namely the intrapsychic states have to be communicated. Emotion theory holds that there is no efficient communication and no profound social relationship without emotional signals being taken into account [61]. Displaying our emotional state on the outside as well as recognizing what other people feel, which constitutes the input and output channels of the affective human-human interaction, might play the main role in an individual's acclimatization in its environment.

In terms of human-human interaction Hess addressed two points of view of what emotions could be [61]. The first point, which goes back in history to the famous book of Darwin [35], states that the displaying of emotions is an innate symptom of the underlying emotional state. This notion is supported by a few recent studies, which presented that specific facial expressions can be linked to specific affective states [44, 127].

According to the second point of view, the displaying of emotion is thought not to provide valid information regarding the underlying emotion, yet serves purely communicative functions. Concerning on the possible difference between the emotions and their expressions Fernandez-Dols and Ruiz-Belda showed that people who have just won a medal tend to show facial expressions different to those commonly associated with happiness even though they tend to report having been happy. This finding seems to suggest that

smiling is not necessarily a sign of a pleasurable experience but rather a social signal [48].

In his above mentioned work, Hess defended that emotion expressions are neither innate symptoms of the underlying emotional states nor serve purely as communicative functions. He stressed the ability of adopting the organon model, which is equally well suited for emotion communication as the original goal it was introduced for. the organon model, which was originally introduced by Karl Bühler for describing the linguistical human-human communication [19], distinguishes between three functions of a message during a conversation, namely, the symbolic, the symptomatic, and the appeal function. The first refers to the sign content of the message and conveys information directed at the interaction partner. The second, the symptomatic function, corresponds to a readout of the individual's state. And the third function concerns the possible action of the interaction partner. For example, *"the expression of sadness signals that the sender experienced an irreversible loss. It also suggests a specific internal state of the sender, characterized by a specific subjective feeling state, as well as by a number of physiological and behavioral concomitants. Finally, it may serve an appeal function by motivating the observer to help or to comfort."* [61].

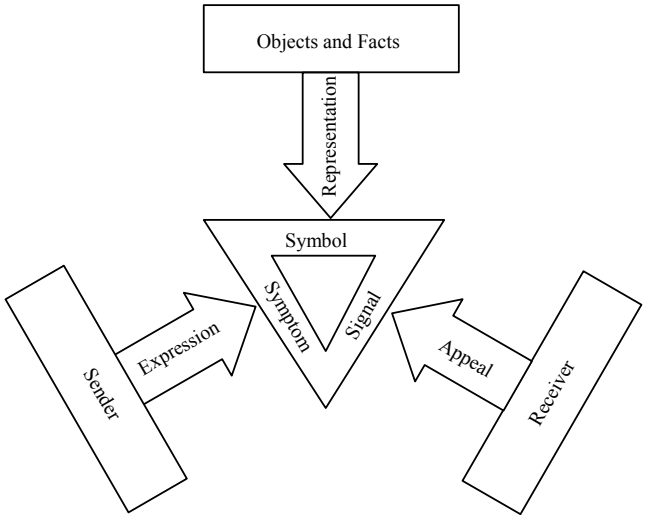


Figure 2.1: *The structure of Emotion according to the organon model proposed by Bühler. The figure is extracted from the work of Hess [61].*

2.2 Emotion Categorization

Diversities of definitions of what emotions are and which organismic processes induce them has led to diversities of models categorizing them. Inspired by Darwinian theory many researchers have been concerned with the categorization of so-called basic or primary emotions. Emotions, however, are assumed, from another point of view, not to be discrete phenomena but rather continuous ones. Psychologists adopting this way of understanding the emotions represent the emotional states in an n-dimensional space (generally two or three). Appraisal theory states that, (I)- emotions are elicited by a cognitive evaluation (appraisal) of an event or situation and (II)- patterning of the possible reactions

(emotion is one of them) is determined by the outcome of this evaluation. The following paragraphs will provide a further detailed overview of these models and how each of them describes the emotions.

2.2.1 Discrete Emotions

Since the publishing of the famous work by Darwin [35], in which he categorized some emotions which he called primary and tried to link each of them to its emotional process useful for survival, many researchers have concerned themselves with the categorization of these primary, i.e. basic, emotions. Basic emotions theories claim the existence of historically evolved basic emotions which are universal and can therefore be found in all cultures. Ortony and Turner reviewed 14 different theories of discrete emotion modeling [103]. Among these theories the number of basic emotions varies from somewhere between two basic emotions in which anger and pleasure are considered the two basic emotions [98], happiness and sadness [163], to 11 basic emotions [2] or even 18 basic ones [51]. Table 2.1 gives an overview of some studies defending the discrete nature of emotion expression[2].

Proposed By	Included Emotions
Arnold [2]	anger, aversion, courage, dejection, desire despair, fear, hate, hope, love, sadness
Ekman et.al. [42]	anger, disgust, fear, joy, sadness, surprise
Izard [67]	anger, contempt, disgust, distress, fear guilt, interest, joy, shame, surprise
Mowrer [98]	pain, pleasure
Otaley et.al. [100]	anger, disgust, anxiety, happiness, sadness
Plutchik [115]	acceptance, anger, anticipation, disgust joy, fear, sadness, surprise
Tomkins [150]	anger, interest, contempt, distress, disgust fear, happiness, shame, surprise
Weiner and Graham [163]	happiness, sadness

Table 2.1: List of studies supporting the discrete nature of emotions. Extracted from the work of Ortony and Turner [103]

Plutcnik preferred to present his model with eight basic emotions by a wheel analogous to the well-known color wheel, as depicted in Fig. 2.2.(a). In Plutchik's model the eight basic emotions are presented in opposite pairs (anger vs. fear, anticipation vs. surprise, trust vs. disgust, joy vs. sadness) on this wheel. The distance of the position of each emotion from the center of the wheel models the activation of the corresponding emotion.

[2] More basic-emotion-based theories are referred in: http://changingminds.org/explanations/emotions/basic emotions.htm

These eight emotions are considered by him to be the *"primary"* ones, from which any other emotion is derived by specific combination *"exp: contempt = disgust and anger, alarm = fear and surprise, etc ..."*. He located them in an arrangement suggesting that the nearer together categories are the similar they are, and the nearer a category is to the center of the circle the more intensity it has *"e.g., low intensity of fear yields timidity while high intensity of it yields terror"* [115].

The most famous and widely accepted approach on basic emotions was conducted by Ekman [42], in which he assumed anger, disgust, fear, happiness, sadness and surprise to be the basic emotions. According to this study each emotion is categorized upon its association with one of the facial expressions, referred as *"prototypes"*, Fig. 2.2.(b) shows examples of these prototypes. These prototypes are assumed to be universally experienced and recognized (independently from sex, age, and culture).

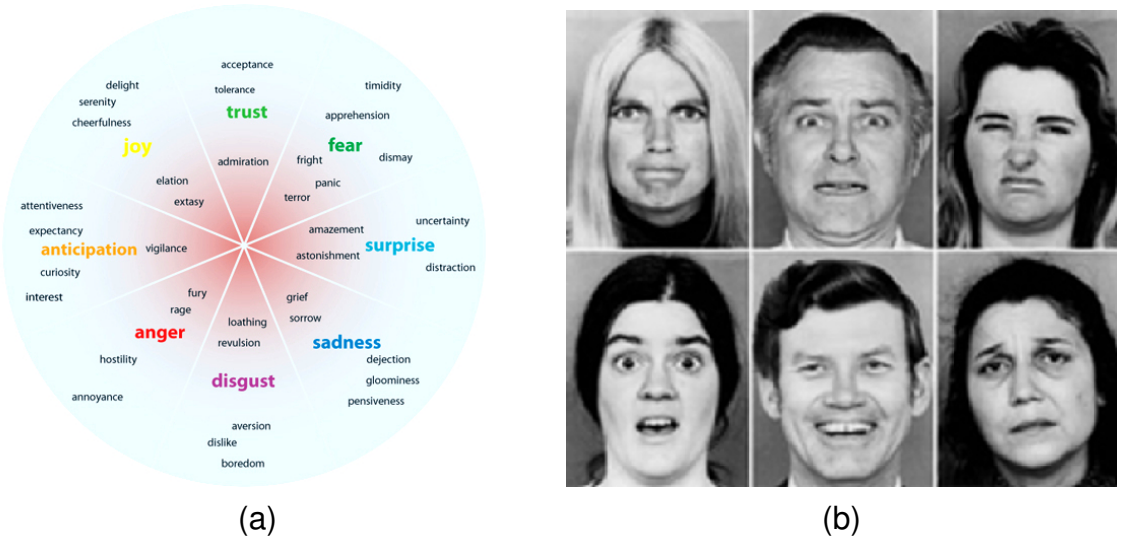


Figure 2.2: *Two well-known theories of basic emotion family. (a) Color-wheel-like location of the eight primary emotions proposed by Plutchik [115], and (b) six prototypes presenting the six basic emotions according to Ekman [42]. From left to right and top to bottom: anger, fear, disgust, surprise, happiness, and sadness.*

2.2.2 Dimensional Models of Emotions

In dimensional theory it is assumed that the basic emotions can be placed in a continuous multidimensional space, in which each dimension stands for a fundamental property common to all emotions. The dimensional model of emotions is closely connected to the semantic differential research method. Employing this method, the raters describe different verbal stimuli on bipolar scales consisting of two opposite adjective pairs, exp., hot-cold, white-black, fast-slow, etc. A study by Mehrabian and Russel provided a great deal of evidence that peoples' ratings of differences in affective meaning can be described by only

three basic dimensions (pleasure, arousal, and dominance) [95]. In this study, they used 18 bipolar adjective pairs, each one rated along a 9-point scale. Applying the statistical factor analysis method on the ratings values of each object, event, or situation, which is wanted to be described, generated the desired scores on the above mentioned three dimensions. Now, the most often used three-dimensional model of emotion is the one of Bradley and Lang. In this model they termed the dimensions as: valence, which ranges from negative to positive emotion, arousal, which ranges from calm to highly aroused, and dominance, which describes if the person is controlled by or controlling the emotion [15]. Figure 2.3.(b) shows some emotions located on both two and three dimensional spaces according to the above mentioned models.

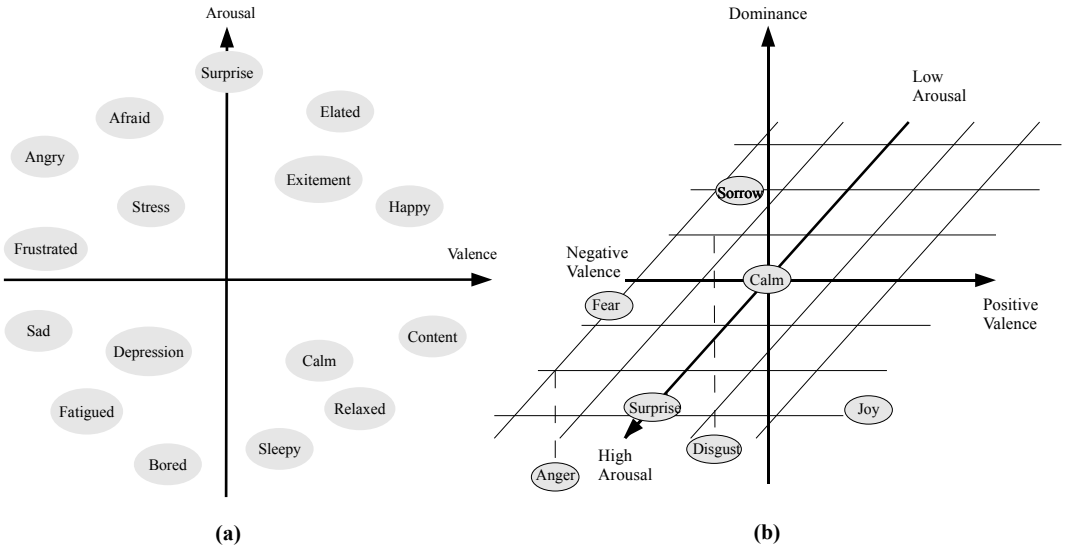


Figure 2.3: *Dimensional models of emotions. (a) two-dimensional valence-arousal judgment space proposed by [124], (b) three-dimensional valence-arousal-dominance judgment space proposed by [15].*

The two-dimensional model of emotions, which was pioneered by Russel and termed the *"the circumplex model"*, might be the most dominant one that is used in the field of automatice recognition of emotions. The model proposed that all affective states arise from two fundamental neurophysiological systems, one related to valence, pleasure-displeasure continuum, while the other is related to arousal. Joy, for example, is conceptualized as an emotional state that is the product of strong activation in the neural systems associated with positive valence or pleasure together with moderate activation in the neural systems associated with arousal [124]. Fig. 2.3.(a) illustrates the locations of some emotions on a two-dimensional space.

2.2.3 Other Models of Emotions

According to the appraisal theory, emotions are elicited by evaluative judgments (appraisal) by an individual of some events of the environment and their implication to him

as well as his own goals and beliefs. Different response components (physiology, emotion expression, action tendencies) present the outcome of this evaluation process.

To describe the relation between emotions and facial expressions, Scherer introduced the term *"Stimulus Evaluation Checks, SECs"*. SECs describe how an organism evaluates stimuli in a sequence of appraisal checks, and how the outcome of these checks results in a specific facial expression. He arranged five stimulus evaluation checks in a fixed sequence, in which the operation of each check depends on the result of the prior one [132].

The first SEC in the sequence is the novelty check which can be broken down into a set of subchecks. This check evaluates whether there is any change in the pattern of external or internal stimulation and if such a stimulation requires attention or not. The next step in the stimulus evaluation sequence is the intrinsic pleasure check. This check evaluates whether a stimulus event is pleasant or not. A pleasant stimulus will induce approach tendencies while an unpleasant stimulus will lead to avoidance or withdrawal. The third check is the goal/need significance check which determines to what extent a stimulus or situation endangers an organism's survival and adaptation to a given environment. It also concerns the satisfaction of an organism's needs and the attainment of its goals. The fourth check in the stimulus processing sequence is the coping ability check. This check provides the ability to successfully cope with a stimulus and free the emotion system from control by this stimulus. Finally the norm/self compatibility check is the final check. It evaluates the significance of a particular action in terms of social consequences. Therefore, this check is only needed by organisms living in social groups.

Ortony et.al. were concerned with the cognitive structure and the implications of emotions. They developed a computational emotion model; its widespread name is abbreviated from their names *"OCC"* [102]. The aim of this model is to characterize a range of psychological possibilities of emotions rather than to describe the emotions themselves or the emotion related processes. The OCC model was established originally as a standard model for emotion synthesis by agents or characters.

2.3 Emotion Encoding

Despite the long term debate on the nature of emotion, it is almost agreed that these underlying states are internal processes which take place inside someone's body. Izard suggested that the emotion reactions involve changes in neurophysiological functions of different brain areas, changes in the neuromuscular activity, changes in behavior, and subjective emotional experiences [67]. These changes are merely observable by their projection to the outside in the form of:

(i) Verbal cues as spoken languages

(ii) Non-verbal cues including:

  a) Body language as facial expression and gestures

  b) Acoustic cues as speech prosody

c) Cues related to internal physiological changes such as heart rate, blood pressure, skin conductivity and temperature, and brain activity

Some of these cues are easy to perceive, such as facial expression, gestures, and speech prosody, while the others (blood pressure, heart rate measurement, brain and muscles activities, skin conductance and temperature) are, without additional auxiliary instruments, beyond the direct ability of humans to be perceived.

Expression of emotions differentiates according to the cues used, when the voluntariness of expressing is considered. Emotional human-human interaction can occur voluntarily, such as pronouncing some specific words related to specific emotions (glad, blithe and bright as indicators of happiness), displaying some deliberate facial expressions, or changes in speech prosody. However, for natural human-human as well as human-robot interaction the displaying of emotions by using facial expressions or speech prosody is assumed to occur involuntary [112]. Fig 2.4 illustrates a model that allocates each one of the above mentioned emotion-related measurements in a two-dimensional space, i.e., (I)- It's significance in affective Human-Human interaction and (II)- It's voluntariness.

As depicted in this model, the positions of facial expression and speech prosody reflect, on the one hand, their relative great impact on human-human interaction and, on the other, their relatively low voluntary level in contrast to other non-verbal emotion-related cues. Following this notion, we employed two stand-alone emotion analysis systems to infer emotions conveyed by facial expression and speech prosody, and a bimodal one that makes use of fusing them together.

Encoding the emotion by diversity of media will be overviewed in the following subsections. Subsection 2.3.1 and subsection 2.3.2 will focus on projecting the internal emotional state onto the outside through changes of facial configuration and auditive components of speech respectively. Inferring the emotional state from measurements of some internal psychophysiological changes will be reviewed in section 2.3.4.

2.3.1 Facial Expression

The face is the most essential medium in the social interactions in the real world. People are recognized and characterized by their faces. The face provides rich information on cognitive states, e.g., interest, puzzlement, frustration, or boredom. The face provides information about the individual identity, about his/her age, and essentially it provides the ability of conveying emotions serving the communicative aspects [37].

Emphasizing the role of facial expression in human-human interaction many social studies argued that humans tend to communicate the most affectively "face-to-face". Due to their occurrence in the interactive context, facial expressions are generally considered to be cooperative signaling systems, benefiting both the expression encoder, who would like to be understood, and the decoder, who strives to understand [50].

Many questions, however, of what could be the association between the emotions and their expression via facial changes, whether the person's facial expression reflects merely the emotional state of this person, whether facial expressions are pure social adaptations, whether the emotion expressions are some kind of combination between these

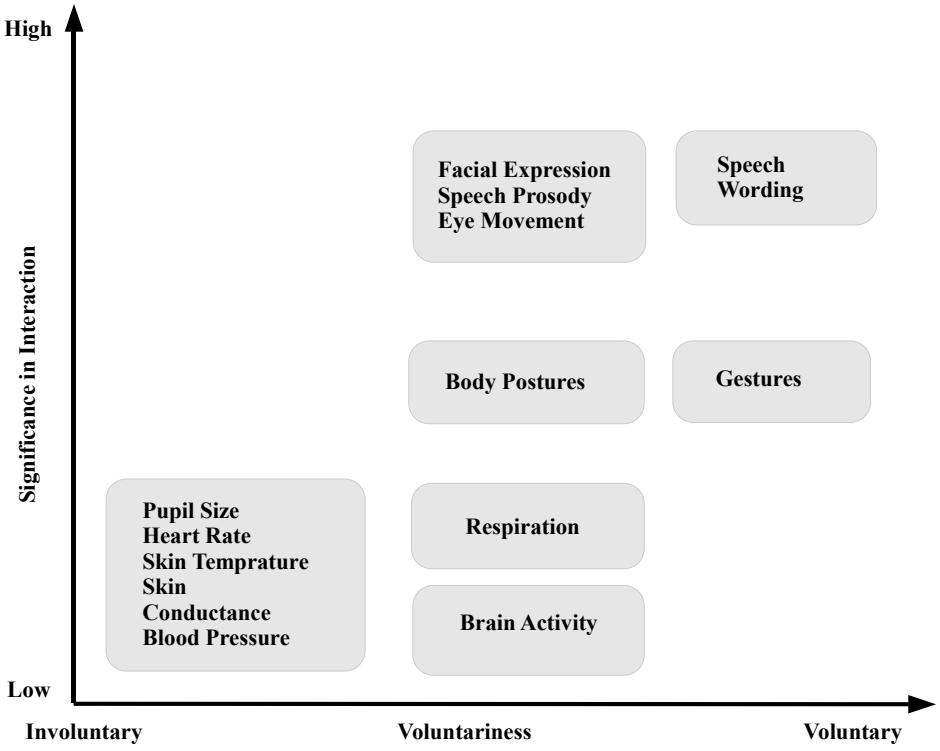


Figure 2.4: *Position of each possible affective measurement according to its significance in affective human-human interaction and voluntariness. Derived from Partala [112].*

two aspects, and many other unanswered questions remain issues of great debate in the psychology of emotion.

Anatomically facial expressions are the outcome of movements of facial skin and connective tissue, which are caused by the contraction of the facial muscles. One of the most noticeable ways to observe a person's emotion is probably through the facial expressions of that individual. The French neurologist and physiologist Duchenne De Boulogne published in 1862 a remarkable treatise on facial expressions. His work was the first to systematically examine the contributions of small groups of cranial muscles to the expressions that communicate the rich experience of human emotion. He reasoned that *"one would be able to paint the expressive lines of the emotions of the soul on the face of man"*. To achieve the goal of understanding how the coordinated contractions of groups of muscles express distinct emotional states he pioneered a so-called, transcutaneous electrical stimulation that used activate single muscles and small groups of muscles in the face, dorsal surface of the head, and neck.

The most prevalent conceptualization of the relationship between the face and emotions is the Facial Expression Program introduced by Russell and Dols [127], which has its roots in Darwin's writings about the face [35]. The key ideas of this model are, (I)- there are a number of basic, universal emotions (6 is an often cited number: anger, disgust,

fear, happiness, sadness and surprise), and (II)- the face reveals the individual's internal emotional state, and though these observers of the face are generally able to read the underlying emotion from the facial expression correctly.

Other researchers adopted that the relationship between experience of emotion and expression is not necessarily relevant [48], while others suggested that the displaying of emotions is manifold of being projecting the inside affective sates into outside and serving as social signals simultaneously [61].

While the above mentioned methods tried to infer what underlines the experienced emotion, many other sign-based coding systems are proposed to describe the association between the expressed facial movements and the experienced emotions by describing the appearance of this emotion rather than its accurate meaning. A diverse mix of such describing models have recently been introduced, e.g., the "Maximally Descriptive Facial Movement Coding System",*"MAX"* that proposed by [68], and the "Facial Affect Scoring System",*"FAST "*, which developed by Ekman et al. [43].

The most widespread standard for describing facial expression may be the Facial Action Coding System *"FACS"*, which was developed by Ekman and Friesen. In FACS all possible facial movements are divided into 44 facial actions *"referred as AUs"* based on visually observable changes in individual's face, eyes, and neck. FACS assigns each muscle group on the face with a code number and so almost every emotional expression may be precisely described using a small set of numbers [113]. Ekman's work has been quite influential in the computer science field, and this conceptualization of the relationship between emotions and facial expression underlies much research in affective systems as will be seen in several points in the remainder of this thesis.

2.3.2 Speech

Like the visual expression of emotions (facial expression) the acoustic expression of them (linguistic, and paralinguistic messages) has its own long history. Systematic treatises of the topic of emotion expression during speech communication can be found in ancient Greek and Roman writings. Darwin, in parallel to his great contribution of studying emotions and their associated facial expression, underlined in his pioneering monograph the primary significance of the voice as a carrier of affective signals [35].

Recently, numerous researchers have identified the tremendous meaning of speech in human-human communication. Generally, three major questions have been challenged in studies on emotion expression in speech, as summarized by Siegwart and Scherer [139]:

(i) How does an emotional state, with all the accompanying physiological effects of respiration, phonatation, and articulation, manifest itself by systematic changes in the acoustic parameters of the voice. To deal with this question, either recorded samples of speech occurring during real emotional conditions or actor portrayals for different types of emotions are obtained. Systematic analysis of the respective acoustic features is then conducted to determine the nature of the acoustic effects. The results of this type of research demonstrate that there are specific patterns of acoustic cues that characterize particular emotional states [135]

(ii) The ability of the listener to infer the nature of the underlying emotion correctly, based only on the acoustic cues of voice. To elucidate this question, emotional voice samples are presented to listener-judges who are then asked to choose the emotion expressed from a set of emotion categories [134]

(iii) Which cues can the listener use to infer the nature of the expressed emotion from the voice? To answer this question, speech researchers have developed elaborate research designs using procedures of partial masking or filtering of specific cues [135], or the acoustically measured cues have been correlated with listener judgments that have been obtained for the same voice samples [31]

However, speech conveys affective information through explicit (linguistic) messages, and implicit (paralinguistic) messages that reflect the way the words are spoken. Considering the verbal part (linguistic message) only, without regarding the manner in which it was spoken (paralinguistic message), might lead to missing some important aspects of the pertinent utterance and even to misunderstanding the spoken message.

Focusing on the paralinguistic messages that convey affective information, a large body of studies in psychology, psycholinguistics, signal processing, and computer science, provides results on acoustic and prosodic features which can be used to encode affective states of a speaker. Measurements related to the pitch, the duration, the intensity, and the frequency spectrum of an audio signal seem to be the reliable descriptors of the emotions associated with speech. Nevertheless, which set of vocal cues can describe the emotional voice the best is still a debate among the researcher in this domain. Section 5.2 will give a review of the state of art of which cues are employed in automatic emotion analysis systems.

2.3.3 Facial Expression during Speech

Facial expression during social interaction is possibly an honest signal of affiliation, or willingness to reciprocate. Among humans, however, social interaction almost invariably involves speech, and there are unique considerations in the adaptiveness of the relationship between facial expression and speech. Facial expression is coordinated with speech at several levels: (I)- the use of muscles of facial expression to articulate speech sounds, (II)- the contribution of facial expressions to the syntactic structure and the meaning of particular utterances, (III)- graded conversational signals that apply to the overall meaning of speech [39].

In addition to its functions on the encoder side, visual information from the encoder's face can strongly influence speech perception, especially when the auditory information is degraded. In one study, the recognition of auditory sentences in noisy environments is improved from 23% to 65% when the perceivers could also see the face of the interaction partner [145]. The finding of Mehrabian can be considered the best conclusion presenting the effectiveness of spoken communications. He indicated that the linguistic part of a spoken message, that is the actual wording, contributes for only 7% to the effect of the

message as a whole. While the paralinguistic part, which is how text is vocalized, contributes for 38%, and the speaker's facial expression contributes for 55% of the effect of the spoken message [94]. This implies that facial expressions form the major modality in emotional human-human interaction during speech too, and have to be considered by HRI systems applicable in conversational sessions.

2.3.4 Internal Physiological Changes

In addition to the affective measures discussed above, there are also quite a few other means for measuring affect-related information from the emotion encoder. The following paragraphs will give an overview of measures suggested to be related to emotions.

- Blood Pressure. Blood pressure is a measure of the pressure at which the blood flows through the body. Happiness, anger, sadness, and anxiety increase blood pressure to differing degrees. Picard suggested that the blood pressure increases with negative emotions such as fear and anxiety, and decreases with relaxation [114].

- Brain Activity. Using an electroencephalograph "EEG", with electrodes attached to the scalp the electrical activity of the brain can be measured. It has been found that EEG asymmetries over the frontal cortex during emotions related to the behavioral tendency of approach (joy, interest, and anger) are relatively greater in the left prefrontal cortex than the right prefrontal cortex. Correspondingly, it has been suggested that during emotions related to the behavioral tendencies of withdrawal (sadness, fear and disgust) EEG asymmetries are relatively greater in the right prefrontal cortex, even though the effect is somewhat less clear [25].

- Galvanic Skin Response. Skin conductance is one common measure in psychophysiology and reflects the ability of the skin to conduct electrical current. Skin conductance has been shown to correlate with autonomic nervous system arousal so that an increase in affective arousal causes an increase in skin conductance [26]

- Heart Rate. This is a commonly used psychophysiological measure related to autonomic nervous system activity and it has been used in emotion research for a long time. Heart rate is generally thought to discriminate between positive and negative emotional reactions. It has been shown to decelerate in response to visual and auditory emotional stimulation. The deceleration is stronger when exposed to unpleasant stimuli than when exposed to pleasant stimuli [14].

- Muscle activations. It is also possible to use electromyography (EMG) to measure the emotion-related activations of either facial muscles or even other body muscles. By using affective picture stimuli many studies suggested that negative experiences are associated with high activity of *"corrugator supercilii"*, while low activity of *"corrugator supercilii"* is associated with positive experiences [80]. Based on the activation of non-face muscles Healey distinguished between low and high stress

levels of a driver by measuring the upper back tension from the trapezium muscle [59].

- Skin temperature. Changes in skin temperature can accompany the occurrence of emotions. McFarland concluded that the negative emotionally aroused states *"e.g., anger"* perpetuated decreases in skin temperature, while the calmer and more positive emotional states perpetuated skin temperature increases [93].

- Respiration. Emotional arousal is associated with faster and deeper respiration compared to rest and relaxation, which are associated with slower and shallower respiration. Utilizing a set of derivatives of respiration signals, Healey successfully distinguished seven self-induced emotions (anger, hate, grief, love, romantic love, joy, reverence) and no emotion [59].

- Tactile information. There are many interaction techniques, which use the tactile modality in human-computer interaction in both input and output, e.g., touch screens and force-feedback output devices. The number of tactile techniques designed especially for affective interaction still, however, small. Qi et al. developed the so-called Pressure mouse. It looks like a normal computer mouse except it is equipped with eight tactile sensors, which measure the pressure with which the user is touching the mouse. Based on pressure information they distinguished between user frustration and non-frustration during computer usage [116].

2.4 Accuracy of Decoding Other's Emotion

In order to evaluate an automatic emotion analysis system reliably, the performance of the human observer should be considered as a reference point. Indeed, many aspects influence this judging process by humans as well as by automatic systems.

Decoding an emotion depends on the affective states of both the encoder and the decoder, the context surrounding the encoding of experienced emotion, and essentially on medium via which the relevant cues are to be encoded and decoded. Table 2.2 depicts the decoding accuracy results of some emotion studies, regarding only the cues that used for decoding and encoding them.

Evidence that is provided in the Table 2.2, suggests that the recognition accuracy for facial expressions outperforms all those of other modality. On the whole, in reviewing the evidence from the studies to date, it can be generally concluded that the recognition of emotion from standardized voice samples attains between 55% and 65% accuracy, about five to six times higher than what would be expected by chance. While in facial expression studies it is generally reported that the emotions are recognized with an accuracy on average of about 75% [133]. Emotion analysis based on some physiological measurements seems to perform comparable to the analysis based on facial expression and speech information cues. Because of their nature of being invasive, which derogates the naturality of interaction, such measurements should be avoided in applications of natural human-robot interaction.

Study	Medium	Selected Categories	Average	Chance	Comments
Scherer et al. [133]	Facial Expression	Six Basic	78%	16.7%	Western Countries
	Facial Expression	Six Basic	65%	16.7%	Western Countries
	Vocal Cues	Neutral & Six Basic but Surprise	62%	14.3%	Non-Western Countries
	Vocal Cues	Neutral & Six Basic but Disgust, Surprise	52%	20%	Non-Western Countries
Banse and Scherer [4]	Vocal Cues	fourteen classes	55%	7.14%	-
Qi et al. [116]	Pressure Mouse	Frustration & Non-Frustration	88%	50%	physical connection
Healey [59]	Respiration	Anger, Hate, Grief, Love Romantic Love, Joy, Reverence	81%	14.3%	physical connection
	Upper Back Tension From the Trapezium Muscle	Low Stress, High Stress	70%	50%	physical connection
Tahakashi [146]	Electroencephalograms, EEG	Joy, Anger, Sadness Fear, and Relax	43%	20%	physical connection
Russel [125]	Facial Expression	Six Basic	84.4%	16.7%	Western Literal
	Facial Expression	Six Basic	72.3%	16.7%	Non-Western Literal

Table 2.2: Accuracy of emotion decoding, according to medium of transmission and the cues used for decoding

2.5 Summary

In this chapter we attempted to answer some primary questions on emotion theory, which serve as the theoretical background to our work. Some widespread definitions of what emotion could be are listed in the first part. Nevertheless, it seems that no clear definition exist that attains consensus. However, which of the above mentioned definitions has its own evidence to substantiate against the others is not for us to judge. All we need for our work is that, firstly the emotional behavior plays a major role in human-human interaction, and secondly, this behavior is presented and observed merely via some specific internal or behavioral changes of the individual reflected into the outside.

The second part of this chapter discussed the three basic theories of emotions that dominate in the psychology research area. The pro-basic-emotion theory researchers advocated the existence of a small number of emotions that universally experienced and perceived. Any other emotion according to them is a kind of combination of these basic ones. Dimensional theory argued that the emotions can be categorized in terms of a small number of dimensions (two, valence and arousal or three, by adding dominance). Appraisal theory focuses on the processes that are involved when experiencing an emotion. It provides through a set of variables a sophisticated elaboration of the causes of emotions, e.g., cognitive evaluations of events, criteria, or checks, rather than directly focusing on the emotions themselves.

When the translation of this knowledge from the psychology research field into emotions-sensitive automatic systems is intended, each one of the above mentioned schemes brings along its own cons and pros. The modern appraisal theories in general, and the cognitive one (OCC model) in particular are suitable for automatic systems that

have the ability of displaying emotions rather than recognizing them, which is beyond the scope of our current work in this thesis.

A drawback of the dimensional based theories is that the projection of the affective state into a rudimentary 2D, or 3D space will lead to some degree of loss in information. Additionally, dimensional clustering of emotions is not intuitive when compared with the discrete one, as will be seen shortly later.

The scheme of clustering the emotions into a small number of basic classes has great benefits over the other schemes because, firstly, using a categorial scheme to judge the emotion displayed by others matches the experience of the ordinary human being in daily life; secondly, the limited number of variables to cope with leads to a remarkable simplicity in contrast to other models (appraisal); and thirdly, the fundamental basics of intuition, universality in displaying, and perception, provides for these systems the possibility to be employed independent of sex, culture, ethnicity, and age.

Following this evaluative conclusion we decided to employ the discrete model of emotion in building systems with the ability to recognize the emotions experienced via multi-sensory media. Explicitly we adopted the well-known and the widespread model of Ekman [42] with six basic emotions in addition to the neutral one.

In the third part of this chapter, the expression of emotion via several media is discussed. Results of some studies presented in section 2.3.4 suggest that some physiological measurements might be used successfully for human-computer interaction, when compared to the usage of the traditional emotional-related cues of human-human interaction (facial expression and speech information). These measurements, however, lie beyond the ability of humans to be directly acquired. They need special equipment to be accurately acquired and processed. This equipment might demand fixed physical connections to the user, which make them unsuitable for natural and human-human-like human-robot interaction.

The last part of this chapter discussed the accuracy of decoding another's emotions. Normally, humans express their emotions and receive other's emotions via several communication media, facial expression, speech information, internal physiological changes,...etc. The results, listed in Table 2.2, suggest that the use of facial expression and speech information is the convenient way toward natural human-human communication and consequently human-robot interaction. This notation is supported by a large body of evidence from studies in psychology, linguistic, and computer science [4, 114, 125, 134, 171].

In real-life situation, however, neither facial expression nor speech information are used separately for emotional human-human interaction. Humans encode and decode emotions via several channels simultaneously. The mutual influence between cues of several channels should be taken into account when aiming at a multimodal emotional human-human interaction. For instance, the observer have to combine the facial expression and the speech information cues to be able to judge if the speaker is smiling or just saying "cheese". Challenging this point, we focus in this work on realizing an emotion analysis system that acts like a human observer by combining both information sources, facial expression and vocal cues, in order to act like a human in the emotional bimodal human-

human interaction.

3 Emotion In Human-Robot Interaction

The technical revolution transformed computers into a daily necessity, not just for the engineers who designed them, but also for almost everyone. Evolving the computer from an orientation towards specialists into an instrument for use by ordinary individuals, from an accessory device to a daily need, from a quite rational computing box to a fully social and interactive attendant, has been aimed at since the 1980s.

Unlike to the traditional human-computer interaction, each individual term of the recently aimed at human-computer interaction has its own terminology. The user should not have to be a computer systems expert; he could be a novice user from any age group, any gender, from one of several cultural backgrounds or diverse levels of education, or even handicapped. Computer does not mean explicitly a desktop computer with monitor, mouse, and keyboard. It could range from the well-known desktop computer to a large-scale computer system, process control system, embedded system, wearable system, or even a robot. Interaction is not constrained to the traditionally known keyboard, mouse, and monitor, but rather the recent interaction channels are expanded to serve for more efficient and affective human-computer interaction (speech, touch, gaze direction, virtual-reality).

Social human-computer interaction has gained the interest of quite a few researchers recently. Nass et al. [99] proposed in their work the paradigm of (Computers Are Social Actors *"CASA"*. In their paradigm, they presented through five empirical experiments that the users interact with the computer in a fundamentally social manner. This social response, according to them, is not caused by the thought of computers being human or human-like, or by the belief that the users are interacting with the programmers who designed them, but rather these social interactions are natural responses to social situations, which are easy to generate, commonplace, and incurable.

Like in human-human interaction, emotions should play an essential role in social human-robot interaction. Brave and Nass emphasized that any interface that does not consider the affective state of the user or fails to manifest the appropriate emotion can dramatically impede the performance, making it untrustworthy and incompetent [17]. Acting on the assumption of emotional influence on human-human communication, the researchers in human-robot interaction field argued that the affective states should be incorporated with the aim to achieve systems that interact with humans like in human-human interaction [114].

Picard defined affective computing as *"computing that relates to, arises from and deliberately influences emotion"* and published the first book in this area [114]. Since then, interest in exploring the role of emotion across a number of subdisciplines within computer science has emerged. In her above-mentioned book, Picard summarized the following possible abilities, which a computing system or a robot can possess:

- Systems that have the ability of emotion recognition
- Systems with the ability of emotion expression
- Emotional intelligent systems, and
- Systems that behave emotionally

By focusing on what the user needs, affective computing systems or robots should fulfill some requirements to constitute the desired emotional human-computer interaction, as depicted in a basic framework of an emotion analysis system illustrated in Fig. 3.1. These abilities are: (I)- sensing, recognizing and understanding the user's affective state, and (III)- adapting to the user affect and reacting appropriately. These two components share a third one that includes the cognitive architecture of the robot as well as an affective model (affective profile) of the user(II) [66].

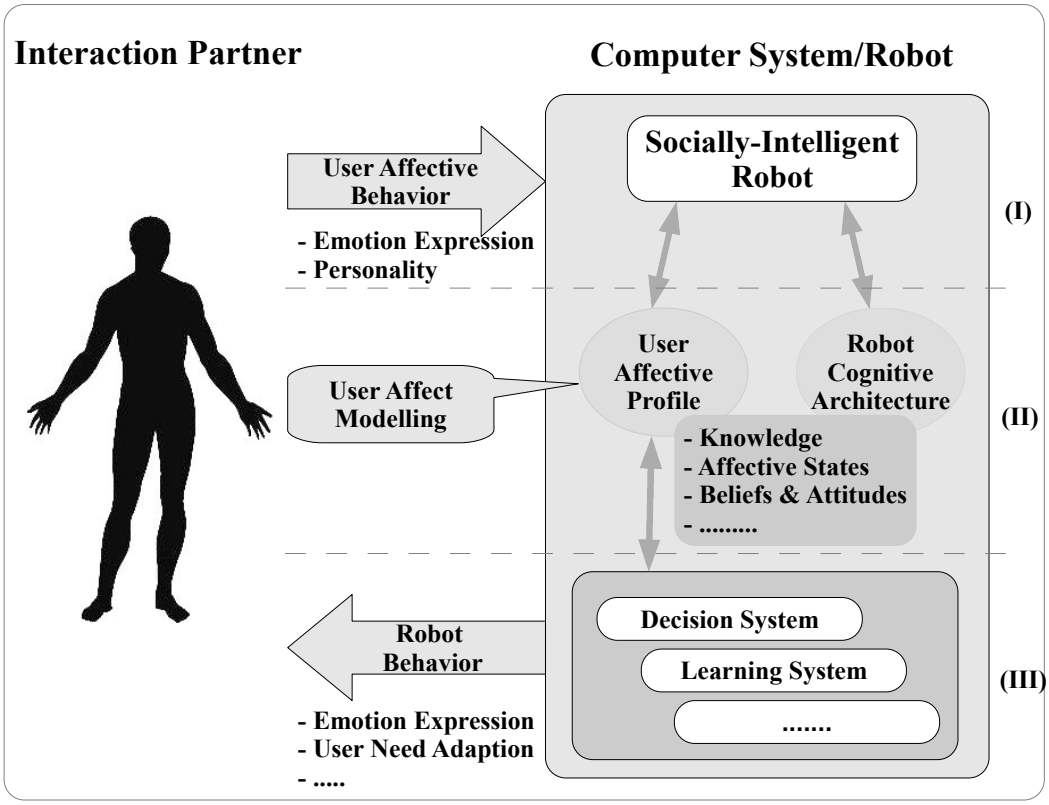


Figure 3.1: *Basic architecture of an Affective Computing framework. The figure depicts the three basic components of such a framework; (I)- Sensing the affective state of the user, (III)- adapting according either to the need of the user or to the cognitive architecture of the system*[1] *and, (II)- Modeling affective behavior of both the user (user affective profile) and the system (cognitive architecture). Derived from Hudlicka [66].*

[1] In our case, the system means an affective system that implemented in a robot

Sensing and recognizing the user's affective state is, indeed, the core aspect of affective computing. Diversity of options already exists for emotion recognition, including not only the traditional cues used in human-human interaction, such as self-reports, facial expression, voice and body language, but rather cues beyond the ability of human to observe such as physiological measurements, as seen in Sec. 2.3.4.

Once the user affective state is identified, a decision needs to be made as to how, or whether, to adapt the system functionality to this state and the art of adapting. For human-computer interaction expressing the affective state of the system is one type of several possible system adaptions, and may be the most appropriate one.

However, the focus of our work in the remainder of this thesis will be on sensing and recognizing the user's emotions − component (I) in the Fig. 3.1 − suggesting it to serve as a basis for further work of both modeling the emotional profile of the user and providing the ability for the robot to react according to the sensed emotion, exp., user imitation in a social human-robot interaction [60, 151].

Acting on, on one hand, the fact that a human perceives another's affective states via multimodality and, on the other, that sensing each modality can influence (increasingly or decreasingly) the sensing of others, as discussed in Sec. 2.3.3, the need for a multimodal affective recognition system is an important issue, when a reliable emotion analysis system is intended.

While unimodal systems (mainly based on facial expression or speech analysis) are investigated deeply, studies taking into account the multimodal nature of the affective communication process are still not comparable. During the last few years, however, numerous researchers have started to examine the scheme of multisensory fusion of two or more modalities. A few attempts and application domains will be listed for illustrative purposes rather than an exhaustive list being given, in order to provide an overview of which effort is achieved and for which field such multimodality affective system can be employed.

3.1 Affective Computing in Use

Since the computer has conquered almost all fields of life, researchers have become concerned with engaging it in human life more efficiently, effectively, and, recently, more affectively. The main goal of affective computing is creating systems with one, or maybe all of the above mentioned affective abilities [114], aiming to employ them in a wide spectrum of daily human life, ranging from application in health-care services [82, 84], in industry [69], in mediating human-human interaction [159], in education [73], games and entertainment [89], and as life companion [18, 144].

3.1.1 Robots with Social Abilities

A very important aspect in developing affective computing systems, and perhaps the most fascinating one, is the research on the integration of such systems into agents or robots with social skills mirroring those of humans. Overall, the use of emotion recognition and

mimicry of the robot is found to be encouraging for further research in a robotic platform for multimodal human-robot interaction [60, 151].

Currently, several humanoid robots are being used in the research field of human-robot interaction, Fig. 3.2 shows some examples of such social robots. Matsusaka et al. developed a torso robot *"ROBITA"* which can participate in a group conversation by estimating who the next speaker will be by speech recognition and face direction recognition [92]. The robot is equipped with video cameras that enable it to detect the gaze and the gesture of its interaction partner. Its face detector is based on quite simple skin color detection and for estimation of gaze direction an eigenface-based classifier is exploited.

"Kismet" is an expressive anthropomorphic robot, which was developed by Breazeal. The robot can engage people in natural and expressive face-to-face interaction. Kismet is equipped with a total of four CCD cameras to visually perceive the person with whom it is interacting. Furthermore, Kismet perceives a variety of natural social cues from visual and auditory channels, and delivers social signals to the human through gaze direction, facial expression, body posture, and vocal babbles. Upon its own emotion model, which combines both discrete and dimensional emotions theories, Kismet can display one of eight affective states (contentment, sadness, anger, fear, acceptance, surprise, sternness, and disgust) when its emotional state oversteps specific criteria [18].

Figure 3.2: *Examples of robots with social skills in research, from upper left to lower right: Leonardo, Kismet, Flobi, and Barthoc.*

Spexard et al. [144] presented an anthropomorphic robot framework *"BARTHOC"* bringing together different interaction concepts and perception capabilities with the goal of creating an interdisciplinary research platform for multimodal human-robot interaction (HRI). The robot uses two cameras and two microphones for capturing visual and audio

information from the robot's surrounding. It has components for face detection, a person tracking module based on anchoring, and extended interaction capabilities based on both verbal and nonverbal communication. *"BARTHOC"* can recognize affect by classifying the prosody of an utterance to seven emotional states (happiness, anger, fear, sadness, surprise, disgust, and boredom) independently from the content in the emotional states of the speaker. The robot is thus able to realize when a communication partner is getting angry and can react accordingly by displaying a calming facial expression on its face. The appropriate facial expression can be invoked from different modules of the overall system, e.g., *"BARTHOC"* starts smiling when it is greeted by a human and stares at an object presented to it. Furthermore, *"BARTHOC"* can mirror the classified prosody of the utterances during the reading of the story (*"Little Red Riding Hood"*), through emotion mimicry of the interactants' facial expression at the end of each sentence they spoke; the expressions were grouped into happiness, fear, and neutrality. As the neutral expression was also the base expression, a short head movement toward the reader was generated as a feedback for non-emotional classified utterances [60].

An embodied computational platform called *"Leonardo"* is implemented by Thomaz et al. Leonardo is an anthropomorphic robot with 65 degrees of freedom that has been specifically designed for expressive social interaction with humans. The robot is able to interact and communicate with people through speech, vocal tone, gestures, facial expressions, and simple object manipulations. The robot has both visual and acoustic inputs [147]. While nothing about the performance of the visual-based affective analysis system is discussed, the acoustic-based system reported as performing well on mapping the encountered affective states in the poor two-dimensional model of emotion (namely, valence and arousal).

In our workgroup[2] an anthropomorphic robot is developed and called *"Flobi"*. The head is suggested to address both sensor head and social interaction requirements. On the sensor side, *"Flobi"* is equipped with a wide-angle, high-resolution stereo camera as visual input channel, and stereo microphones for speaker localization and speech recognition. When it comes to the sociality the exterior of the head is designed in such a way that it affects the interactant in a positive way. It has eighteen degrees-of-freedom (DoF): 3 in the eyes, 2 in the eyebrows, 4 in the eye-lids, 3 in the neck, and 6 in the mouth, in addition to two LEDs (one red, one white) that are placed behind each cheek to enable it to display emotion realted facial expression in a human-like way [86].

3.1.2 BIRON, Social Interactive Robot

In our work in this thesis we focus on embedding a bimodal emotion analysis system (combination of facial expression and speech information) in a robot called **BIRON**, Fig. 3.3 depicts its basic hardware components. The focus of both unimodal systems as well the bimodal one will be on recognizing six Ekmanian basic emotions in addition to the neutral one. The proposed systems fulfils most of the requirements of an ideal emotion analysis system of being fully automatic, real-life applicable, and having the ability

[2] Applied informatics, Bielefeld university, Germany

to deal with affective interaction situations, in which the interaction partner is engaged in a conversational session (i.e. displaying facial expression and speaking simultaneously rather than consequently). Furthermore, the proposed system has the ability to be applied in scenarios of affective interaction with multiple users.

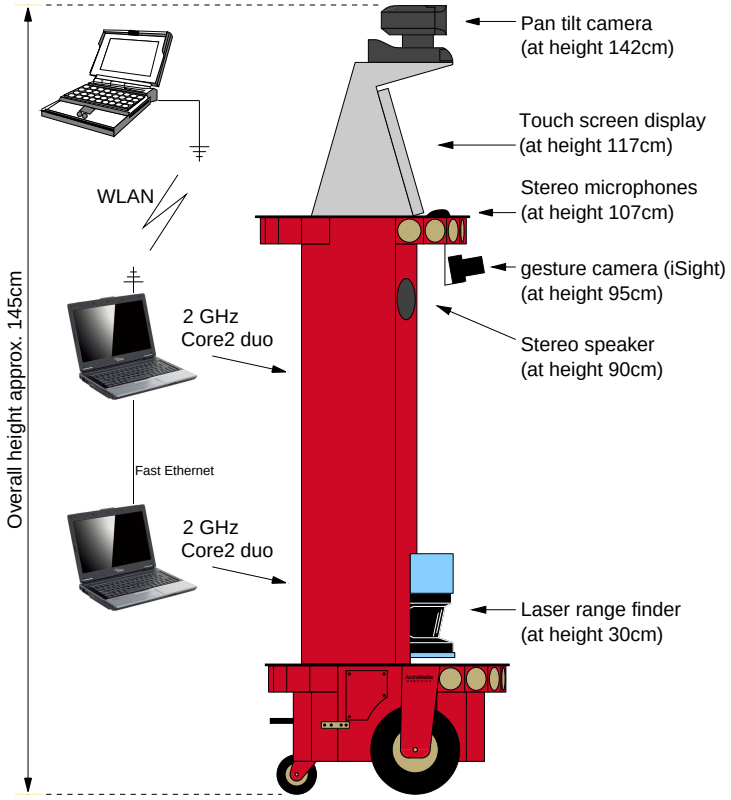


Figure 3.3: *Physical Characteristics of the robot, BIRON..*

BIRON, *"**BI**lefeld **RO**bot companI**ON**"*, is a mobile robot equipped with a series of sensors enabling it to perform well in real-world human-robot interaction scenarios [53]. It is comprised mainly of two Duel Core Notebooks running the Linux operating system, that provide the computational power and to enable the robot to run autonomously in real human-computer interaction. These two laptops serve for controlling the motors as well for socially interaction skills. A laser range finder (SICK LMS200) is integrated inside the front side of the robot base. This laser scanner has the ability to sense objects, for instance the legs of an interaction partner, up to 50m from the robot. For acquiring images of the upper body in general, and the face of the interaction partner in particular, a pan-tilt camera (Sony EVI-D31) is mounted on top of BIRON with approximate height of 142 cm. The camera has the ability to scan an area of ± 100 degree in front of the robot in order to detect possible objects in its field of view. To localizing the interaction partner according to the direction of the acquired sound signal the robot is equipped with two interfacial microphones which are mounted on the upper front side. Two stereo speakers serve as output medium of the speech dialog, for whose input channel serve the above mentioned microphones. Furthermore, BIRON possesses some social abilities, such as detecting and identifying its interaction partner, and adapting to some of his/her verbal cues.

3.2 Enabling Affective Interaction with BIRON

The focus of this work is to provide the ability for the robot to behave emotionally in real-life human-robot interaction scenarios. More precisely, we provide the ability of emotion understanding of the interactant for BIRON, according to which the robot can adapt its behaviour. As discussed so far, facial expression and speech information are the most honest ways of conveying the internal emotional state on the outside. Following that inferring the interactants' emotions will be based on analyzing cues provided by these channels. Emotion recognition based on facial expression analysis will be discussed in depth in chapter. 4. A fully automatic, online-applicable system will be proposed that will enable BIRON to label the facial expression of its interaction partner with one of seven basic classes of emotions. Adapting to the emotional state of the interactant is an open issue for future work.

Situations, in which emotions are communicated by displaying facial expressions in a deliberative way, are seldom encountered in natural human-human as well as human-robot interaction conditions. Usually the display of facial expressions is associated with speaking. During such sessions facial configurations related to the process of speech production can conflict with those facial configurations associated with the experiencing of emotions. Speech production can include lip movements, mouth configurations and the movements of the lower part of the face. Altogether these configurations can lead to a derogated performance of the facial-expression-based emotion analysis system.

Humans exploit cues of all available modalities associated with the experience of emotions. Thereby, not only consent results of different modalities lead to more confident decisions, but conflicting results can also be helpful. Multimodal treatment can help in detecting falsified or masked emotions, or finding out more reliable modalities for certain emotions. Regarding translating this notion to our robot, taking speech information into account would smooth the effect of speaking on inferring the emotion displayed by facial expression during speech. To achieve that, a probabilistic-based fusion model is proposed in chapter. 5. The model makes use of combining both modalities according to the respective discrimination power.

4 Facial Expression Analysis for HRI

The expression of emotion used to be primarily a research subject of psychologists, neuro, and social scientists, see chapter. 2. However, the rapid technical and industrial evolution invalidated this restriction. Since the computer has conquered our lives and became an essential need rather than an accessory, more-sophisticated human-computer interaction is being aimed to which will enable users to interact with computers more socially and effectively.

It is widely accepted that the face plays a major role in human-human communication. The face is a rich source of information from an individual. The face conveys information about individual identity, social identity and character, the individual's internal emotion states, and gaze [37]. A large evidence gained from the theoretical studies of emotion suggests that humans communicate more effectively face-to-face, i.e., facial expression of emotion should play the major role in this interaction [61]. On other side, the rules of human-human interaction have to be obeyed if the aim is to achieve reliable human-computer interaction [114, 121]. Hence automatic recognition of facial expressions should be considered when the development of natural human-robot interfaces is aimed at.

In general the human face conveys information via four kinds of signals. Pantic and Bartlett [107] categorized them into four basic classes and concluded that only signals of the last category can be employed for both encoding and decoding an individual's emotions.

- **Static facial signals:** represent the permanent structure of face features, such as the bony structure, the soft tissue, and the overall proportions of the face. These signals contribute to an individual's appearance and are usually exploited for person identification.

- **Slow facial signals:** represent changes in the appearance of the face that occur gradually over time, such as development of permanent wrinkles and changes in skin texture. These signals can be used for assessing the age of an individual.

- **Artificial signals:** are exogenous features of the face such as glasses and cosmetics. These signals provide additional information that can be used for gender recognition.

- **Rapid facial signals:** represent temporal changes in neuromuscular activity that may lead to visually detectable changes in facial appearance. including blushing and tears. These atomic facial signals underlie facial expressions.

Indeed, automatic analysis of rapid facial signals seems to play a main role in various vision systems, including automated tools for tracking gaze and focus of attention, lip

reading, bimodal speech processing, face-based command issuing, and emotion-related facial expressions. When natural, social, and emotional interaction between humans and computers, in general, and especially robots is aimed at, facial expressions provide the most convenient way to communicate basic information about needs and demands to the machine. Internal emotional states of the interaction partner such as happiness, indicated by smiling in Fig. 4.1, can often be read from the displayed facial expression and then trigger an appropriate dialog and adaptive behavior in the robot.

Figure 4.1: *Facial expression in interaction between human and robot. Interactant's smile can be understood as an acceptance of what the robot has done and trigger a suitable behavior.*

The initial focus of automatic facial expression analysis was on the recognition of the prototypical emotions from posed static input. Early attempts focused on recognizing prototypical emotions from two static face images *"neutral and expressive"* [97, 129]. In the second half of the 1990s, automated face expression analysis started focusing on posed video sequences and exploiting temporal information in the displayed face expressions [13, 45].

New findings from neuroscience, psychology, cognitive and computer science, on the one hand, and the rapid revolution in computer skills, on the other, inspired numerous researchers to embark on building machines with more-sophisticated skills of facial expression analysis. Recently, the analyzing of facial expression, which occur in real-life human-robot interaction, has gained the interest of quite a few researchers.

When human-human-like human-robot interaction (via facial expression) is aimed at, the human visual system should be set as a reference point about the desired functionality of the employed systems. To achieve that, the employed facial expression analysis system should fulfill many requirements:

- **The first requirement** is that all of the stages of the facial expression analysis, namely, face detection, facial expression information/features extraction, and facial expression classification, have to be performed automatically
- **The second requirement** is that all the above-mentioned processes have to be performed in real-time conditions. Otherwise, delayed reaction of such systems renders the interaction desynchronized and less efficient, and

(a) (b)

Figure 4.2: Special constrained conditions of facial expression analysis systems. (a) Two cameras are mounted to an arrangement that keeps the head fixed in order to enable the system to capture specific views of the face. (b) The face is labeled with colored markers; they are tracked to infer some emotion-related displacements.

- **The third requirement** is that an ideal facial expression analysis system should like the human visual system by being robust in inferring the affective state of the interaction partner in most real-life situations

For these situations the exploited system ought to be applicable in everyday-life conditions, able to deal with rigid head motions, changes in lighting conditions, changes in viewing conditions, and partial occlusion of the face. Additionally, an ideal system should have the ability of inferring the affective state of the interaction partner regardless of sex, age, and ethnic group. Furthermore, it has to perform well without assuming the observed subjects to have any constrained appearance (colored regions on the face, or face markers [71], without it being set in any constrained environment (human sitting fixed in front of a camera that is mounted on the subject's head and placed in front of his/her face [110], or without being invasive (the user is connected physically to the system) [16, 25, 73]. Fig. 4.2 illustrates some of these restrictions that may derogate naturality of the interaction.

In this chapter, a system for visual facial analysis that fulfils almost all of the above-mentioned requirements will be presented. The proposed system is fully automatic and exhibits noticeable robustness in every day-life conditions with either a single interaction partner (person-dependent) or multiple partners (person-independent). The system is also embedded as part of the interactive robot companion **BIRON**, which serves for natural human-robot interaction in real-life scenarios.

The state of the art in the field of automatic facial expression analysis will be overviewed first. In this overview an analytical discussion about the three basic stages of an automatic facial expression analysis system-, (namely face detection, facial features extraction, and classification) will be included. The focus of the overview will be on the sufficiency of these methods to be applied in real-life scenarios.

A novel approach to fully automatic facial expression analysis with the ability to be

applied in real-life scenario will be discussed in Sec. 4.6. The system performs fully automatically ranging from the first stage of face acquisition to extracting the proper facial features to the categorization of these features into a suitable emotion class. An integration concept of this system in our BIRON as well as a comprehensive evaluation on a suitable database (DaFEx) will be discussed in Sec. 6.3. The performance of this system in natural and unconstrained human-robot interaction will be discussed in Sec. 6.6.

4.1 Related Work

In a related work that discusses systems with online ability, Valsatar and Pantic [152] reported some progress of building a system that enables fully automated fast and robust facial expression recognition from face video. They analyzed subtle changes in facial expression by recognizing facial muscle action units (AUs) and analyzing their temporal behavior. Their work was based upon a set of spatio-temporal features calculated from tracking data for 20 facial fiducial points. To detect these 20 points of interest in the first frame of an input face video, they utilized a fully automatic, facial point localization method that uses individual feature GentleBoost templates built from Gabor wavelet features. To track the facial points they employed a particle filtering scheme that uses factorized likelihoods and a novel observation model that combines a rigid and a morphological model. The AUs displayed in the input video and their temporal segments are recognized finally by Support Vector Machines trained on a subset of highly informative spatio-temporal features selected by AdaBoost. However, like all geometric-based methods, which will be discussed shortly later, their work demands labeling of the first frame as reference. Furthermore no online ability is discussed. Bartlett et.al. [7] proposed a system that automatically detects frontal faces in the video stream and classifies them into seven classes: neutral, anger, disgust, fear, happiness, sadness, and surprise. The face detector, which is based on Viola & Jones detector, is used to convey an image patch containing the face to a Gabor-wavelet-based facial feature extractor. Gabor representation of the conveyed patch is formed and processed by a bank of SVM classifiers.

4.2 Structure of Face analysis system

Fig. 4.3 illustrates the basic components of the system of face analysis. The first task in the face analysis process is to find the faces in the input image/frame, if there are any. Depending on the application, the faces may be tracked over time or detected from a single image (or from a video image, but without tracking). To avoid undesired effects (changes in light conditions, face pose or scale, ...) many preprocessing methods can be employed (normalization, face patch segmentation, ...). The next stage is to extract some emotion-related information of the detected/tracked facial image/frame. The final step is to categorize the image/frame either in one from a set of basic emotion classes or by labeling them with a single action unit or a combination of several action units according to the extracted information from the previous stage.

Recent vision-based affective computing systems can be categorized according to several aspects: the information to be processed, how to process them, and what the expected output of such systems is. Facial information processing may occur holistically (the whole face) or locally (areas prone to change with facial expressions, such as brows, mouth and lips). Some methods (motion-based approaches) focus directly on facial changes associated with emotion-related facial expressions, while others (deformation-based approaches) rely on natural face images as reference in order to extract facial features relevant to facial expression. The output of such systems can either infer what underlies the displayed expression (emotion) or solely describe specific movements on the face surface.

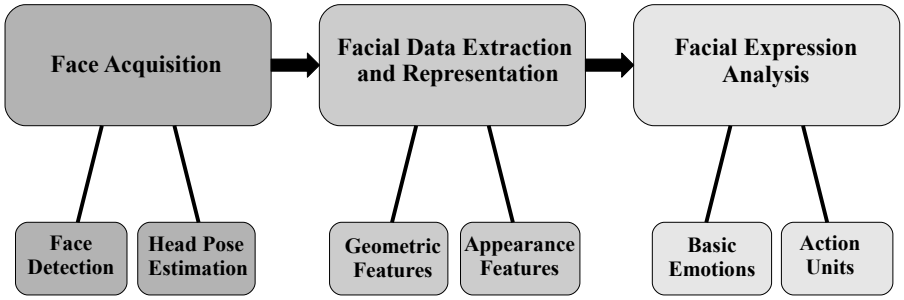


Figure 4.3: *Schematic architecture of the facial-expression-based emotion analysis system. The first stage serves for capturing input images, and finding or estimating the location of the face in these images. The middle stage uses the information provided from the first stage to extract some facial features related to the displaying of emotion. These features are finally labeled with one of a predefined number of basic emotions or action units in the facial expression analysis stage.*

Section 4.3 will give an overview of face detection approaches that have recently been used for face analysis, and then discuss the ability of each to be applied in real-life scenarios. Quite a few methods are applied in order to extract some specific feature suitable for the next stage of classification/analysis. The common methods employed by the computer vision community for this task will be discussed in Sec. 4.4. Classification based on the extracted features constitutes the last stage of an automatic expression analysis system. This phase includes categorizing the displayed facial expressions into one of a predefined number of classes (basic emotion), action units (AUs), or even combinations of both. Some approaches in the field of pattern recognition and machine learning, which currently dominate in the field of automatic face analysis, will be overviewed in Sec. 4.5.

4.3 Face Detection

Face detection is the first step in any automatic face analysis system. Its quality determines the quality of the following stages. Face detection means finding the face in an input image, if there is any. Face detection is a challenging task since there are many conditions that may vary. When the applications in real-life conditions are considered,

the pose and the orientation of the face in relation to the camera can vary temporally; the target face may be partially occluded by some other objects, or by another face; and the image including the face may be taken outdoors in daylight, indoors in fluorescent light, or in other lighting conditions. However, when the analysis of facial expression is aimed at, the variation in the individual's face should be taken into account. Each person has a unique face, which looks different, has its own biometric, and displays facial configuration differently. Furthermore the face of the same individual looks different when the time of image acquisition is considered (age of the person, eyeglasses, beard, moustache and make-up make, ...).

4.3.1 Basic Approaches to Face Detection

Yang et al. [166] classified the used approach into four major categories, namely *" knowledge-based approaches, feature-based approaches, template matching approaches, and appearance-based approaches"*, under which further approaches are categorized according to the method used.

The hierarchical top-down knowledge-based approach assumes a different face model at different coarse-to-fine scales. For efficiency, the image is searched at the coarsest scale first. Once a match is found, the image is searched at the next-finer scale until the finest scale is reached [165]. Another aspect of knowledge-based face detection is employed to detect some facial features: eyes, nose, and mouth. In this approach, the facial features are located using both horizontal and vertical projection of the image intensity. The local minima of the horizontal profile correspond to the left and right boundary of the face, while those of the horizontal profile determine the locations of lips, nose tip, and eyes [78]. However, projection methods suffer from two drawbacks, namely when the face has to be detected from an image with complex background and the case of multiple faces.

The bottom-up feature-based approaches search through the image for a set of invariant facial features and groups them into face candidates based on their geometric relationship. Quite a few methods are used to detect faces via detecting some features such as face contour, eyebrows, eyes, nose, mouth, hair line, or combinations of several features. Some examples are: Edge map *"Canny Detector"* to segment the face from a cluttered background [142], and searching the points and the edges from an image then attempting to group them together[167].

The neural networks displayed their efficiency in solving several pattern recognition problems such as autonomous robot driving and object recognition, to which the two class pattern recognition problem of face detection belongs. In neural-networks-based face detection techniques, the structure of the network is chosen first. The structure defines how many layers the network will have, the size of each layer, the number of inputs of the network and the value of the output for faces and non-faces. Then the network is trained using samples of faces and non-faces. To test an input image for faces, most of the approaches apply a window scanning technique to detect faces. The window has a fixed pre-determined size and moves with certain step until it has scanned all parts of

the input image. Each time the output is computed if it is above a certain threshold the window is classified as face. The most common way to give the ability of detecting faces with different scales is the forming of an image pyramid by successively resizing the input image. Then each level in the image pyramid is scanned by the moving window. Incorporating this approach with a multilayer neural network yielded a successful neural-network-based face detection system, which is proposed by Rowley et al [123].

According to the deformable templates, the facial features are described by a parameterized template. Snakes or active contours are commonly used to detect a head boundary. The evolution of a snake is achieved by minimizing an energy function by utilizing an optimization technique [57]. Images of human faces lie in a subspace, which can be represented by several statistical analysis methods, an example is the using of component analysis, *"PCA"*. Sirovich and Kirby [76] proposed a technique using principal component analysis *"PCA"* to represent human faces. The technique first finds the principal components of the distribution of faces. Each face in the set can then be approximated by a linear combination of the largest eigenvectors, more commonly referred to as *"eigenfaces"*.

The AdaBoost-based face detector, introduced by Viola and Jones, demonstrated that faces can be fairly reliably detected in real-time under partial occlusion. The achievements of theirs can be attributed to the fast-calculated Haar-like features via the integral image and the cascade structure of classifiers learned by AdaBoost [155]. The motivation behind the cascade of classifiers was that simple classifiers at an early stage can filter out the most negative examples efficiently, and stronger classifiers at a later stage are only necessary for dealing with instances that look like faces.

4.3.2 Selected Face Detection approach

To evaluate the performance of a face detector several metrics can be considered, such as detection accuracy, detection speed, required training time, the required number of training samples, the sensitivity of the environment, and the memory requirements during training and test. All these biometrics are crucial when it comes to applying facial expression analysis systems for natural human-computer interaction.

The detection rate and false alarm rate are typically used to determine the detection accuracy. The detection rate can be defined as the ratio between the number of correctly detected faces and the number of faces in the image. The false alarm rate on the other hand determines the number of detected faces that are not actually faces.

The detection speed is usually an important factor, especially when real life applications are being aimed at. There are huge differences in detection speeds between detection methods. The methods that work in real-time with a standard PC are the most useful in typical HRI applications since users usually expect immediate feedback on their actions. Some examples of such face detection methods are the cascaded face detector by Viola and Jones [155], the rotation invariant multi-view face detector by Huang et al. [63], and the Encara face detector proposed by Castrillón et.al. [22].

Another metric to be considered is how precisely the face has to be located so that the

detection is considered correct. If the application is interested only in the number of faces in the image or just the rough location is needed, then all the detections for the face can be considered correct. However, if further classification has to be done for the detected face (in our case, the extracting of features related to facial expression) then badly located faces may become a problem even if face alignment is used.

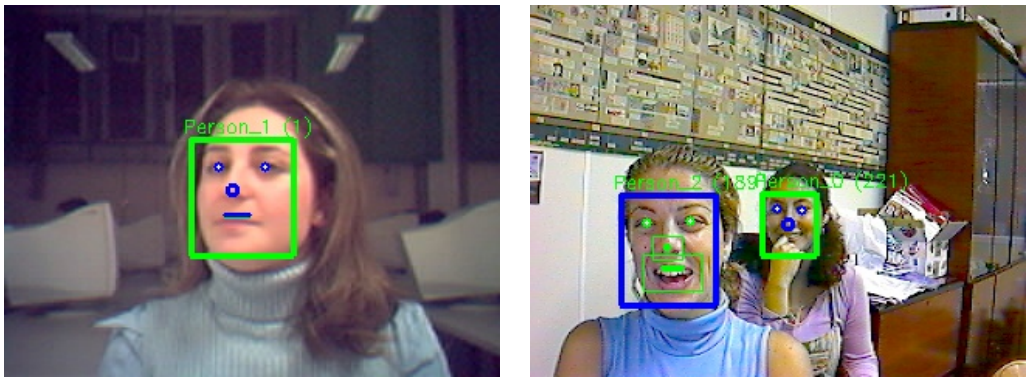


Figure 4.4: *Examples presenting the performance of the exploited face detector in single-user and multiple-user situations; (a) and (b) respectively.The bounding box of the face and the positions of the eyes, nose, and mouth are colored with either green or blue. Green indicates that the detected object is a frontal face, while blue indicates the using of tracking rather than simple detecting. Printed by courtesy of Castrillón [22].*

Considering the above mentioned metrics, we employed the face detection approach proposed by Castrillón. This approach makes use of a combination of the two feature- and knowledge-based families. Such a combination enables ENCARA to inherit the robustness of the former family and the detection speed of the latter. In addition to the position of the face, ENCARA delivers information about the positions of the eyes, nose, and mouth, which can serve as initialization points for the facial features extractor, see Sec. 4.6 for more details. ENCARA achieves approximately 99.9% correct detection on faces and 87,5% on eye pairs in real-time data with multiple faces enabling the applicability in the scenarios of emotional interaction with multiple users, as illustrated in Fig. 4.4 with an approximate speed of 20-30 msec on recent PC hardware [22].

4.4 Facial Feature Extraction

After the face location has been detected, the next step of the automatic facial expression analysis system is the extraction of some facial features relevant to facial expression. Fruitful avenues of feature extraction methods can be found in the literature [149, 171]. Current used methods can be categorized according to their method of answering three major questions [107]:

(i) Is temporal information being used?

(ii) Are the features holistic (the whole face) or local (subparts of the face)?

(iii) Are the features view- or volume-based (2D or 3D)?

Given the goal of face recognition (description of facial behavior looking at an individual's face), most of the proposed approaches are based on two-dimensional facial features, either for the whole face or just some facial regions.

The features extracted using the recently proposed 2D-based methods can be categorized into three main classes, namely geometric features, appearance features, and hybrid features. The features of the first class are obtained using geometrical information (motion) of either the whole face shape or some facial features (the shape of some facial components, eyes, mouth, or the location of some fiducial points, corner of the eyes, mouth, eyelids) [28, 149]. The features of the second class represent the texture of the facial skin including wrinkles, bulges and furrows in the whole face [83] or around some facial features [5]. Hybrid features are presented as a combination of the above two methods [29, 164]. It is suggested and almost agreed that using both geometric and appearance features may be the best choice for presenting information relevant to facial expression [107, 149, 171].

4.4.1 Geometric-Based Facial Feature Extraction

Optical flow approaches gained their own place in the field of describing facial features' motion. That is because the dense flow information is available throughout the entire facial area, regardless of the existence of facial components, even in the areas of smooth texture such as the cheeks and the forehead. Many researchers adopted this approach to capture the information associated with geometric displacements either locally (specific set of facial regions) [104] or holistically (the whole face) [81].

Lien [81] analyzed holistic face motion with the aid of wavelet-based, multi-resolution dense optical flow. For a compacter representation of the resulting flow fields they computed PCA-based eigenflows both in horizontal and vertical directions. Otsuka and Ohya [104] estimated facial motion in local regions surrounding the eyes and the mouth. Feature vectors are obtained by taking 2D Fourier transforms of the vertical and horizontal optical flow fields.

Some researcher have adopted this approach more recently [1, 149]. Some limitations, however, are inherent in optical flow techniques, such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and essentially the changes in illumination.

In feature point tracking, the estimates of geometric displacements are obtained only for a selected set of prominent facial features such as lines and furrows in the regions surrounding the eyes and mouth. In order to reduce the risk of tracking loss, feature points are placed into areas of high contrast, preferably around intransigent facial features (eyes, eyebrows, mouth). Hence, the movement and deformation of the latter can be measured by tracking the displacement of the corresponding feature points. However, as facial motion is extracted at only selected feature point locations, other facial activities are ignored altogether. The automatic initialization of feature points is difficult and often done manually [161].

It is possible to determine facial deformations with more reliability than with previously discussed methods, namely by measuring deformation in areas where there is underlying muscles contraction or retraction. These are mostly skin regions with relatively poor texture. Highlighting for these methods is required and can be done by either applying color to salient facial features and skin or by affixing colored plastic dots or colored markers to predefined locations on the subject's face [70, 96]. It is to be noted, however, that the tracking of feature points or markers allows facial-expression-related motion/deformation information to be extracted from often only specific feature point locations, while the rest is neglected. Furthermore, labeling the face with colored markers will lead to such methods not fitting a natural human-human-like human-robot interaction.

Yet another way of extracting the motion is difference images. When it comes to the analysis of facial expression, difference images are mostly created by subtracting a given facial image (exp, image with a sad face) from a reference image, containing a neutral face of the same individual [38]. In comparison to optical flow approaches, no flow direction has to be extracted, but only differences of image intensities. However, accurate face normalization procedures are necessary in order to align reference faces to the test faces.

Similar to difference images methods, Valsatar and Pantic proposed an approach based on temporal templates. Temporal templates are 2D images, constructed from image sequences, which show motion history; that is, where and when motion in the image sequence has occurred [153]. In addition to the sensitivity to rotation, translation and scale changes, temporal templates rely on the assumption that either the camera and the background have to be static or the motion of the object of interest (the face) has to be well separable from the motion induced by both camera movements and background clutter.

Black and Yacoob introduced local parametric motion models that not allow only non-rigid facial motions to be accurately modeled, but also provide a concise description of the motion associated with the edges of the mouth, nose, eyelids and eyebrows in terms of a small number of parameters. However, the employed motion models are focused on some specific facial regions involved in facial expressions (eyes, eye-brows and mouth), while the analysis of other features, occurring in residual facial areas, was not considered [13].

The works of [27, 138] focused on the design of Bayesian network classifiers for emotion recognition from face videos based on facial features tracked by a method called *"piecewise Bezier volume deformation tracking"*. This tracker employs an explicit 3D wireframe model consisting of 16 surface patches embedded in Bezier volumes. The main shortcoming of these mentioned model-based methods is the demand of manual selection of landmark facial points in the first frame of the input video based on which the face model will be warped to fit the face [27, 138]. For this reason they are not sufficient for the goal of natural and human-like human-computer interaction.

In addition to the fact that each one of the above-mentioned methods has its own drawbacks, all geometric-based methods share the shortcoming of demanding a reference image to be compared with the test images. This problem is solved either by labeling the first image of each sequence with the suitable emotion state or by collecting data in which each sequence has to begin with a specific facial expression, almost neutral. Rather than most of the existing geometric-based methods are not suitable assuming some conditions

(outplane head motions, partial face occlusion). These limitations, indeed, reduce the usability of system based on such approaches to perform well in the application of neutral real-time life-like human computer interaction.

4.4.2 Appearance-Based Facial Feature Extraction

The geometric-based facial feature extractors mentioned above are based on the thought of motion when the problem of facial expression recognition is considered. It is suggested by Bassili [10] that humans can recognize facial expresions above chance from motion using point-light display. In contrast, it is argued that the recognition of facial expression via texture (appearance-based) outperforms those based on motion (geometric-based) [38, 174].

Appearance-based facial feature extractors methods can be subcategorized into Gabor filters, integral-image-filters *"haar filters"* based methods, neural-nets-based methods, and kernel-based approaches including (principle component analysis, PCA, and independent comonent analysis, ICA).

Gabor wavelets are 2D sine waves modulated by a Gaussian envelope. They are widely used to extract the changes in face appearance as a set of multiscale and multiorientation coefficients. Gabor filters may be applied to the whole face image [8, 83] or to specific regions of the face [88, 158].

In the work of Vukadinovic and Pantic, the face is detected using the Viola and Jones detector [155]. The detected face is then divided into 20 relevant regions of interest (ROI), each of which is examined further to predict the location of a corresponding facial point. The proposed facial feature point detection method uses individual feature patch templates, which are built from a combination of gray levels and Gabor wavelet features of the corresponding ROI, in order to detect points in the relevant region of interest [158].

A number of approaches to face image analysis have employed data-driven kernels, which are learned from the statistics of the face image ensemble, to represent expressive facial images. Representations based on principal component analysis, eigenfaces for the whole face region and eigenvectors for some facial features (mouth and eyes regions), are applied successfully to recognize facial expressions [105]. PCA has some advantages over other face recognition schemes in terms of speed and simplicity, and its reduced sensitivity to noise (exp, noise due to small occlusions; as long as the topological structure does not change, changes in background). However, PCA-based methods are not robust against pose changes since global features are highly sensitive to translation and rotation of the face. Furthermore, when handling video stream (real-life applications) it is considered that PCA-based methods offer a high extent of sensitivity against rigid movement of the observed face [175].

4.4.3 Hybrid Methods of Facial Feature Extraction

The importance of appearance-based features for expression recognition is emphasized by several studies, which suggest that appearance-based features may contain more infor-

mation about facial expression than displacements of a set of points [38, 174]. Zhang et.al. compared the two above-mentioned types of facial features extraction method, namely the geometric positions of 34 fiducial points on a face and 612 Gabor wavelet coefficients extracted from the face image at these 34 fiducial points. The recognition rates for seven emotion-specified expressions (the six basic emotions of Ekman plus the neutral one) were significantly higher for Gabor wavelet coefficients [174]. Similarly, Donato et al. compared several techniques utilized to extract facial features for recognizing six single upper-face AUs and six lower-face AUs. These techniques include optical flow, principal component analysis (PCA), independent component analysis (ICA), local feature analysis (LFA), and Gabor wavelet representation. The best performances were obtained using a Gabor wavelet representation and independent component analysis [38]. In contrast to that, Pantic and Patras suggested that the motion/geometric-feature-based methods outperform the appearance-based ones [108].

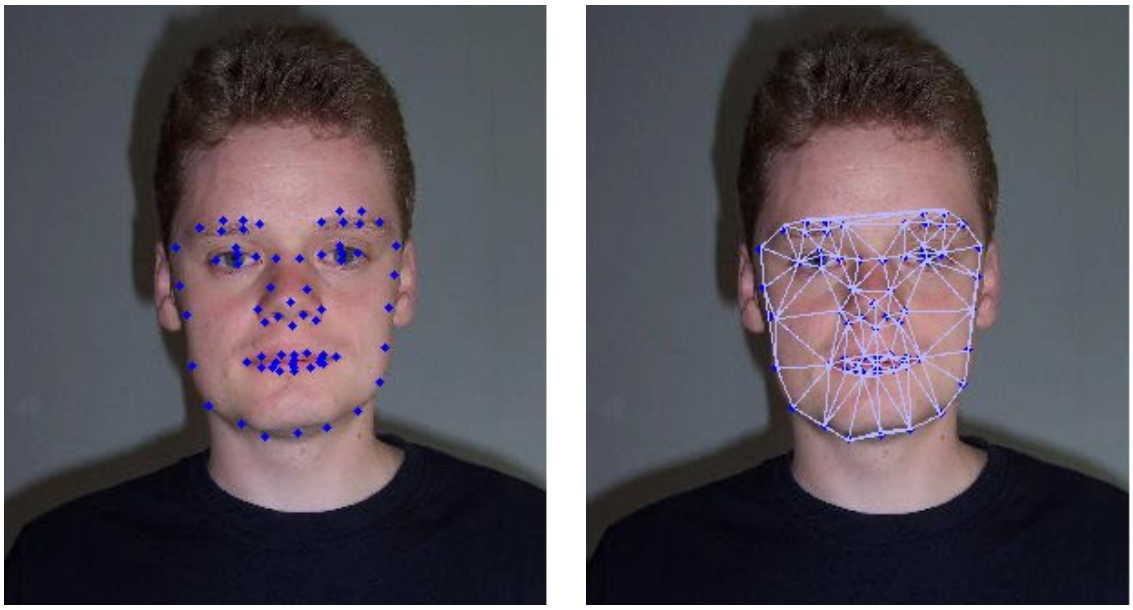
While geometric-feature-based approaches do relay only on the location of a specific set of facial points (either the whole face or some facial features) and the appearance-based approaches present solely the changes in the texture of the face (skin changes), combining these two approaches makes use of elements of both streams with the aim of harnessing their advantages. This notion is supported by evidence provided from studies of the human visual system [10]. Indeed, it can be concluded that combining appearance-based and motion-based representations may be the most powerful method for extracting facial features sufficient for face analysis in the application field of human-computer interaction [6, 108, 171].

Wen and Huang used a motion-based explicit 3D wireframe face model to track geometric facial features defined on the model. A 3D model is fitted to the first frame of the sequence by manually selecting landmark facial features such as corners of the eyes and mouth. Gabor wavelets are then used to extract the appearance changes in 11 facial regions, which are being tracked using the 3D model, as a set of multi-scale and multi-orientation coefficients [164].

Tian et al. studied geometric features as stand-alone an in combination with Gabor filters. To detect and track changes in the shape of some facial components; mouth, eyes, lips, brow, and cheek, they used a multi-state model for each component. To extract appearance-based feature Gabor wavelet coefficient are then calculated in 20 locations which are defined based on the geometric features in the upper face [148]. However, while the region of the face and approximate location of individual face features are detected automatically in the initial frame, the contours of the face features and components have to be adjusted manually in the this frame. Zhang and Ji employed the same appearance-based method, as used by [149], while for some geometric features they used 26 facial points around the eyes, eyebrows, and mouth instead of multi-state models of Tian [173].

4.4.4 Active Appearance Models, Feature Extractor

Active Appearance Models *"AAM"*, which were proposed by [29], are a powerful generative class of methods for modeling and registering deformable objects. AAMs simul-



(a) (b)

Figure 4.5: Training an AAM. (a)- An image example of annotated set for training an AAM, the blue points indicate fiducial landmarks. (b)- Triangulation method used to warp each image in the training set to match the base shape .

taneously model the intrinsic variation in shape and texture of a deformable visual object as a linear combination of basismodes of variation. As such modes of variation can be easily calculated by applying PCA to a normalized training set. The result is a compact model, capable of generating large variations in shape and texture with a relatively small parameter set.

In order to build an AAM, some set of annotated images is needed. The annotation is achieved by putting (drawing) some fiducial landmarks on the images of the training set. The points represent the shape of the observed object; in our case, to annotate a face object the point can be set on the boundaries of the mouth, nose, eyes, chin, eyebrows, the center of the mouth, and the irises. This annotation process can be achieved either manually or by employing a bootstrapping method, in which the model built from the already annotated images serves as an initialization method for annotating new images. This process is iterated until all images in the training set are annotated. AAMs computed from image-data annotated using Bootstrapping-based method, however, seem not to extract feature sufficient for facial expression recognition. Researching for suitable alternative set an open issue for future work. Fig. 4.7(a) shows an image example with suitable annotation.

The shape of an AAM of such an annotated set of images is generated as the following:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s}\mathbf{b_s} \tag{4.1}$$

where $\bar{\mathbf{s}}$ is the mean shape (base shape), $\mathbf{P_s}$ is a set of orthogonal modes of variation in the training set and $\mathbf{b_s}$ is a set of shape parameters. Both $\bar{\mathbf{s}}$ and $\mathbf{P_s}$ are obtained by applying Procrustes alignment algorithm and principal component analysis on the annotated images

of the training set.

The texture of an AAM is defined within the so-called *"shape free"* frame. It consists of N pixels, usually chosen to lie within the convex hull of the base shape $\bar{\mathbf{s}}$. To achieve that, each image in the training set should be warped so that its markers match those of the base shape. For the warping task a triangulation algorithm is used. Fig. 4.7(b) shows the result of applying a triangulation method on the image displayed in Fig. 4.7(a).

As with the shape model, the texture is also generated using a linear combination of basis variation vectors:

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{P_t b_t} \tag{4.2}$$

where $\bar{\mathbf{t}}$ is the vectorized mean image, $\mathbf{P_t}$ is the texture basis matrix that is obtained by applying PCA to a set of images, after being warped according to the shape-free frame and $\mathbf{b_t}$ is the set of grey scale parameters.

In order to obtain more compact representation of deformable objects it is crucial to take the possible correlation between the variations in both shape and grey level in account. To achieve this, a third PCA is applied on the concatenated $\mathbf{b_s}$ and $\mathbf{b_t}$ parameters. After this process has been achieved, each example from the training set can be represented as follows:

$$\mathbf{b} = \mathbf{Qc} \tag{4.3}$$

where $\mathbf{Q}$ are the eigenvectors and $\mathbf{c}$ is the vector of appearance parameters controlling both the shape and the grey scale of the model.

Given a new unseen image $\mathbf{I_i}$, AAM fitting is the process of finding the model parameters which best fit the considered image. In other words, the process is to find the best parameters that minimize the difference between the new image and the one synthesized by the model $\mathbf{I_m}$, i.e.

$$\delta\mathbf{I} = \mathbf{I_i} - \mathbf{I_m} \tag{4.4}$$

This is usually an iterative process which sequentially updates the model parameters p through an update function:

The power of this generative model stems from (I)- its compact representation of appearance (comprising shape and texture), (II)- its rapid fitting to unseen images and, (III)- its robustness in terms of handling variations in image intensity and feature shape. In addition, being PCA-based, AAM is insensitive to partial occlusion and can represent semi-profile faces rather than frontal view [30]. In contrast to other feature extraction methods, AAM needs neither Manual labeling of the first frame in each sequence nor labeling of the emotion class.

AAM, nevertheless, could fail when it comes to the application in real-world scenarios. That is due to the fact that AAM needs an initial estimate of the position, orientation, or scale at which this model should be placed in an image. Considerable face zoom variations or out-plane/rigid movements of the face may cause the fitting algorithm of AAM to fail when the suitable initialization is neglected. Fig. 4.6 illustrates an example of such a

(a) (b)

Figure 4.6: AAM fitting algorithm can fail if rigid head movement, exp., from the position in (a) to the position in (b), is encountered. The images are extracted from the DaFEx database [11].

fitting fail. To overcome this drawback we have proposed a novel initialization method that causes the fitting algorithm of AAM to be employed faster and in a more robust manner in our facial-expression–based emotion analysis system for real-life conditions, see Sec. 4.6 for more details.

4.5 Classification

The last step of an automatic facial expression analysis system is to translate the extracted information (movement information, parameter vector) into a suitable description of the displayed configurations. The currently used facial expression analyzers classify the encountered expression (i.e., the extracted facial changes information) as either a particular facial action (AUs) [110, 152] or a particular basic emotion [3, 117, 137, 169]; some systems perform both [109, 173]. Basic-emotion-based and facial-action-units-based models are referred by Fasel and Lüttin as judgment- and sign-based approaches respectively [47]. While the aim of former is to infer what underlies a displayed facial expression, such as affect or personality, the latter aims solely to describe the surface of what is shown, such as facial movement or facial component shape. As an example, upon seeing a frowning face, an observer with a judgment-based approach would make judgments such as *"angry"* whereas an observer with a sign-based approach would code the face as having activation of some specific AUs.

For sign judgment approaches the most widely used method for manual labeling of facial actions is the Facial Action Coding System. FACS is a human-observer-based system designed to detect subtle changes in facial features viewing videotaped facial behavior in slow motion. Trained human observers can manually code all possible facial displays, referred to as action units AUs. Action units may occur individually, in double or even multiple combinations [113]. Ekman and Friesen proposed that specific of these combinations represents prototypic expression of emotion. Emotion-specified exp-

ressions, however, are not included in standard FACS; they can be coded in separate systems, such as EMFACS or FACSAID [1].

One of the main criticisms of AUs-based works is that the methods are not applicable in real-life situations, where subtle changes in facial expression typify the displayed facial behavior rather than the exaggerated changes that typify posed expressions [170]. Furthermore, the extraction of AUs from faces is a complex process especially if the data is collected in natural environments, the background is likely to be complex, the face may have any size and position or may be partially obscured, and the subject may be engaged in conversational session, in which it is difficult to distinguish between facial changes related to emotion displaying, and those related to speech production processes. All this implies a relatively complex image processing problem that is coupled with the difficulty of interpreting the AUs and relating them to emotional state, yielding a complex multi-disciplinary research problem.

When it comes to the basic-emotion-based approaches, it is suggested and widely accepted that there are some basic emotions, which are universally encoded and decoded, see Sec. 2.2.1 for more details. The most commonly used facial expression descriptors in message judgment approaches are the six basic emotions of Ekman, namely; (angry, disgust, fear, happy, sadness and surprise), and usually accompanied by a neutral one.

This trend can also be found in the field of automatic facial expression analysis. Most researchers in the field of facial expressions analysis developed so far, have targeted human facial affect analysis systems, which attempted to recognize all these basic emotions [3, 117] or a small set of them like neutral, positive and negative emotions [137, 169], or even only distinguishing affective from non-affective faces [168].

Regardless of the categorization scheme used, two basic mechanisms of classification are applied by the expression analyzers; namely, static approaches and dynamic approaches.

4.5.1 Static Approaches

They rely solely on the feature (information) extracted from the current frame to infer the encountered expression. The input image of such classifiers can be a static image or a frame of a sequence that is treated independently. Variety of methods of this family can be found in the literature for facial expression recognition such as Neural Nets, which are either applied directly to the face image [49] or in combination with some facial features extraction methods, such as PCA, ICA, Gabor wavelet, or AAM [105]. Linear Discriminant analysis [160], Bayesian network classifier [27, 70], K-nearest neighbor [138], and Support Vector Macines [96, 117] are further methods, which are used to classify facial expression statically.

In the following, two approaches of this family will be reviewed and discussed regarding their suitability to be applied in an online, real-life-applicable facial expression analysis system. The nearest neighbor classifier is very fast, even for high-dimensional feature vectors, and therefore it is especially suitable for real-time processing. However,

[1] http://www.face-and-emotion.com/dataface/facsaid/description.jsp

it yields slightly lower classification rates compared with the SVM classifier, as will be discussed in Sec. 6.3. Hence, SVM will be primarily employed in this thesis.

Nearest Neighbor Classifier

The k-nearest neighbors (kNN) rule is one of the oldest and simplest methods for pattern classification. Nevertheless, it often yields competitive results in certain domains, when it is cleverly combined with prior knowledge. The kNN rule classifies each unlabeled example by the majority label among its k-nearest neighbors in the training set. Its performance thus depends crucially on the distance metric used to identify nearest neighbors. In the absence of prior knowledge, most kNN classifiers use simple Euclidean distances to measure the dissimilarities between examples represented as vector inputs. Euclidean distance metrics, however, do not capitalize on any statistical regularities in the data that might be estimated from a large training set of labeled examples.

The most common metric often used in face recognition and facial expression analysis is the Mahalanobis distance, which is computed as follows

$$D(x, z) = (x - z)^T (W)(x - z) \tag{4.5}$$

where x and z are observation vectors being compared and (W) is a weighting matrix (covariance matrix of training data). Nearest neighbor classification relies on the assumption that the class conditional probabilities are locally constant. This assumption becomes, however, false in the case of high dimensionality with finite samples set leading to an unsatisfying performance.

Support Vector Machines

Most machine learning algorithms receive input data during a training phase then build a model of the input and deliver a hypothesis function that can be used to predict future data. Among these algorithms, support vector machines (SVMs) pioneered by Vapnik [154], have received considerable attention because of their superior performance in pattern recognition and function regression. In the following, we concisely review the basic principles of SVMs for pattern recognition. In a simple two-classes case, given a set of labelled training pairs such as:

$$\mathcal{S} = (\mathbf{x}_i, y_i), i = 1, \ldots, l \text{ where } \mathbf{x}_i \in \mathbf{R}^n \text{ and } \mathbf{y} \in \{1, -1\}^l \tag{4.6}$$

The main goal of the SVM approach is to define a hyperplane in a high-dimensional feature space $\mathbf{Z}$, which divides the set of samples in the feature space such that all the points with the same label are on the same side of the hyperplane. In general, the mapping from the input space to the feature space is non-linear and can be expressed as $\phi : x \in \mathbf{R}^n \longmapsto \mathbf{z} \in \mathbf{Z}$ Therefore, the training problem of an SVM in this case is to find w and b so that

$$f_{w,b}(\mathbf{z}) = sgn(\mathbf{w}^T \mathbf{z} + b) \tag{4.7}$$

where $\mathbf{w}$ is a coefficient vector and b is a bias of the hyperplane; sgn $\big[\,\cdot\,\big]$ stands for the bipolar sign function. Depending on the kernel used, input data is mapped into either another space with the same number of dimensions, exp., the linear kernel, or usually in a higher-dimensional feature space through some nonlinear mapping chosen a priori so that a nonlinear classification boundary in input space can be achieved.

By making use of the nonlinear mapping ϕ, one can map the set of training samples in input space as a corresponding training set in the feature space, i.e., $\phi : \mathbf{x}_i \longmapsto \mathbf{z}_i, i = 1, ..., \mathbf{N}$. An optimal separating hyperplane (OSH), which maximizes the margin between the two closest vectors to the hyperplanes, is constructed in this feature space. The classifier constructed by the hyperplane in the high-dimensional feature space can be built as

$$y_i\big[\mathbf{w}^T\mathbf{z}_i + b\big] \geq 1, \qquad i = 1, ..., \mathbf{N} \tag{4.8}$$

Among the separating hyperplanes, the one with the maximal distance to the closest point is called the optimal separating hyperplane, which will result in an optimal generalization. In view of the fact that the distance to the closest point is $1/\|\mathbf{w}\|$, two fold of the distance is called the margin.

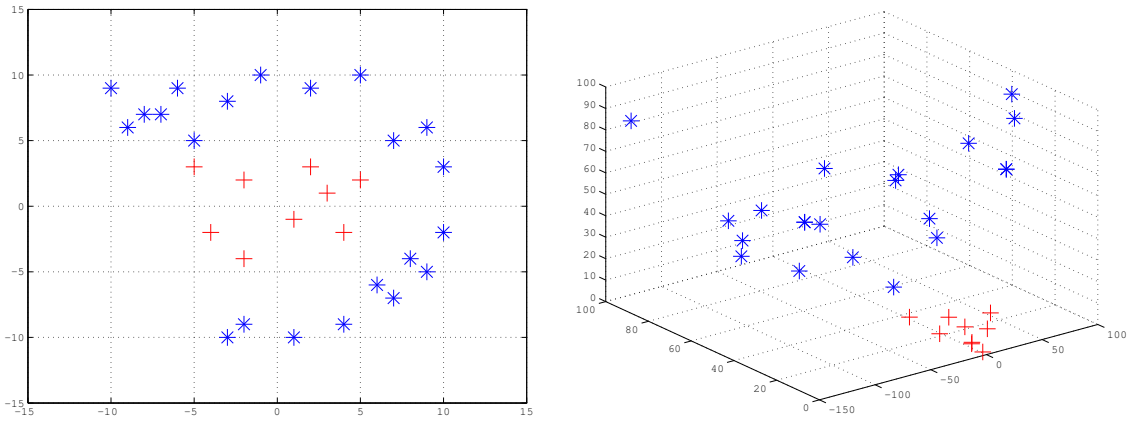
The margin, which is presented as the distance between the optimal separating hyperplane and the closest point, can be seen as a measure of the generalization ability of this hyperplane classification. The larger the margin, the better the generalization is expected to be. Thus, the optimal separating hyperplane is the one which maximizes the margin and gets the best generalization performance.

For the nonseparable case, on the other hand, slack variables ξ_i can be introduced such that the support vector machines require the solution of the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\mathbf{W}^T\mathbf{W} + \mathbf{C}\sum_{i=1}\xi_i \\ subject\ to \quad & y_i\big(\mathbf{W}^T\phi(x_i)\big) \geq 1 - \xi_i, \\ & \xi_i \geq 0 \end{aligned} \tag{4.9}$$

Here training vectors $\mathbf{x}_i$ are mapped into a higher (maybe infinite) dimensional space by the function ϕ; $\mathbf{C} > 0$ is the penalty parameter of the error term. The purpose of these non-negative variables is to allow misclassified points to exist. When the i th example is misclassified by the hyperplane, the corresponding $\xi_i > 1$. As a consequence, $\sum_i \xi_i$, is a measure of an upper bound on the number of training errors with which minimization reduces the empirical training error.

As mentioned above, SVMs perform an implicit embedding of data into a high-dimensional feature space, where linear algebra and geometry may be used to separate data that is only separable with nonlinear rules in input space. To achieve that, the learning algorithm is formulated to make use of kernel functions, allowing efficient computation of



(a) (b)

Figure 4.7: An example of mapping into a high-dimensional space. While the random samples in the two-dimensional space are not linearly separable (a), they can be more easily separated after mapping them in a three-dimensional space (b).

inner products directly in feature space, without the need for explicit embedding. Fig. 4.7 shows an example of mapping from two-dimensional into three-dimensional space. Given a nonlinear mapping ϕ that embeds input vectors into feature space, kernels have the form:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \tag{4.10}$$

In view of the used kernel, SVM will separate the training data in feature space by a hyperplane defined according to the type of kernel function used. However, four types of kernels are generally used with SVM classification applications [23].

- **Linear:** $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$.
- **Polynomial:** $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + t)^d$, with t the intercept, and d the degree of the polynomial
- **Radial Basis Function (RBF):** $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$, with σ the variance of the Gaussian kernel
- **Sigmoid:** $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \tanh(s\mathbf{x}_i^T \mathbf{x}_j + b^2)$, with s the scale parameter and b the bias

The SVM approach is highly modular, allowing domainspecific selection of the kernel function used. In contrast to other learning approaches, SVMs allow for some intuition and human understanding. They deal with noisy data and overfitting by allowing for some misclassifications on the training set. Multi-class classification is accomplished by a cascade of binary classifiers together with a voting scheme (one-against-all). SVMs are successfully employed for quite a few classification tasks in general and for face analysis especially. SVMs currently outperform artificial neural networks in a variety of applications. Their high classification accuracy for small training sets and their generalization performance on data that is highly variable and difficult to separate make SVMs particularly suitable to a real-time approach to expression recognition in video [96].

4.5.2 Dynamic Approaches

They take into account the temporal pattern in displaying facial expression. Hidden Markov Models, which are commonly used in the field of speech recognition [119], have proved their usability for facial expression analysis as they allow the dynamics of facial movements to be modeled. Several HHM-based approaches are applied in this field, especially in combination with geometric-based extraction methods [5, 27]. In order to describe a real-world process, such as variations in facial expression with an HMM, an appropriate selection of HMM parameters is required; this process is known as HMM training. Generally, an HMM is described with the following set of parameters:

$\lambda = (AB\pi)$, where $A = \{a_{ij}\}$ is the state transition probability matrix, $B = b_j(O_t)$, is observation probability distribution, and $\pi = \pi_j$, is the initial state distribution.

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N$$

$P(O_t|q_t = S_j), \leq j \leq N$, and

$$\pi_j = P(q_1 = S_j)$$

In order to use HMM for facial expression recognition, the probability $P(O|\lambda)$ has to be computed, with which HMM will generate an output observable symbol sequence $O = o_1, o_2, ..., o_T$ given the above-mentioned parameters of HMM, with T as the sequence length. Normally, this process is achieved by enumerating each possible state sequence of the whole length T. In this case, it will be N^T possible combinations of state sequence, with N as the number of states. The critical disadvantage of HMM is the computation complexity. Considering the above-mentioned observation sequence, the overall time complexity of computing the probability of $P(O|\lambda)$ is from the order $O(N^T T)$. Calculating this value is very difficult, even if a small number of states and frames are considered.

Cohen et al. compared the performance of several avenues of static classifiers based on Bayesian nets with the performance of a temporal one based on HMM. The conclusion of them work was: *"It seems, both from intuition and from our results, that dynamic classifiers are more suited for systems that are person-dependent due to their higher sensitivity not only to changes in appearance of expressions among different individuals, but also to the differences in temporal patterns. Static classifiers are easier to train and implement, but when used on a continuous video sequence, they can be unreliable especially for frames that are not at the peak of an expression. Another important aspect is that the dynamic classifiers are more complex, therefore they require more training samples and many more parameters to learn compared with the static approach"* [27],P.p, 183.

4.6 Contribution

In the following, we present a novel approach of a fully automatic real-life-applicable emotion recognition system based on analyzing the displayed facial expression. The proposed system fulfils most requirements desired for an ideal system, which are extensive discussed at the beginning of this chapter. The general architecture of the facial analysis system will be presented first. It based on a novel approach of initializing methods of AAM, which enable the system to be applied in a more robust manner and fully automatic in real-life affective human-robot interactionThe effect and relevance of the several initialization types on the performance of the AAM fitting algorithm will be discussed. A comprehensive evaluation on eligible databases that unveils the relevance of the developed initialization schemes will be discussed in Sec. 6.3.

The goal of this facial analysis system as part of the general architecture is to recognize facial expressions that are displayed by the interaction partner of the mobile robot BIRON. Emotion recognition is achieved not only when the user displays one of the six basic emotion purely and deliberately, but also when he/she displays them during speech, which is seldom challenged in this field of research.

In our visual-based emotion recognition system, the core technique for extracting some features related to emotion is AAMs. AAM facial feature extractor is embedded in a vision system that consists of four basic components as illustrated in Fig. 4.8. Face pose and basic facial features (BFFs), such as nose, mouth and eyes, are recognized by the face detection module. The coordinates representing these features are conveyed to the facial feature extraction module. Here, the BFFs are used to initialize the iterative AAM fitting algorithm. Several methods are proposed to initialize the AAM on the basis of detected BFFs. After feature extraction the resulting parameter vector for every image frame is classified into one of the six basic emotions in addition to the neutral one; two classification types will be discussed in Sec 6.3, namely nearest neighbor and support vector machine. Besides the feature vector, AAM fitting also returns a reconstruction error that is applied as a confidence measure to reason about the quality of the fitting.

To ensure the online ability, the system provides a soft real-time applicability that running at a rate of approximate 5Hz on recent PC hardware.

For face detection we preferred to employ an approach that makes use of cue combination, instead of restricting our process to a single image-based technique such as the well known [155], to get greater robustness and higher processing speed, particularly for our scenario where live video is processed. The face is initially detected by means of Viola & Jones- based detectors. This initial detection allows the system to opportunistically trigger the search of its inner facial details: eyes, nose and mouth (which will be called the basic facial features BFF as from now). Some detection results are presented in Fig. 4.9. Further details about the used face detection approach can be found in [22]. Our assumption is that their detection would improve the precision of the initialization and therefore the AAM search process, as will be discussed later. Thus, once the face has been detected, the facial feature detectors are launched in those areas that are coherent with their expected location for a frontal face. Those located will characterize the face as follows:

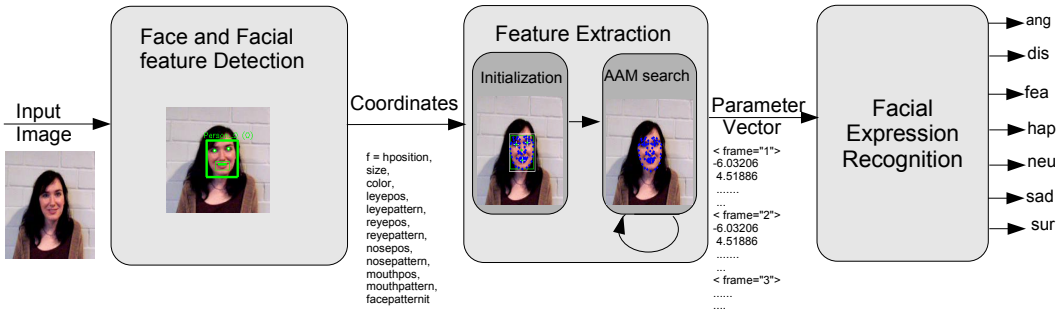


Figure 4.8: Schematic Architecture of the proposed facial analysis system. Positions of the face and some facial features are extracted in the first stage, left. Coordinations of these features are then used to align an AAM, middle. Parameter vectors, which are extracted by AAM, are then classified by a SVM model, right

$$f = \langle position, size, color, \quad leye_{pos}, leye_{pattern}, reye_{pos}, \quad reye_{pattern}, nose_{pos}, nose_{pattern}, mouth_{pos}, mouth_{pattern}, face_{pattern} \rangle$$

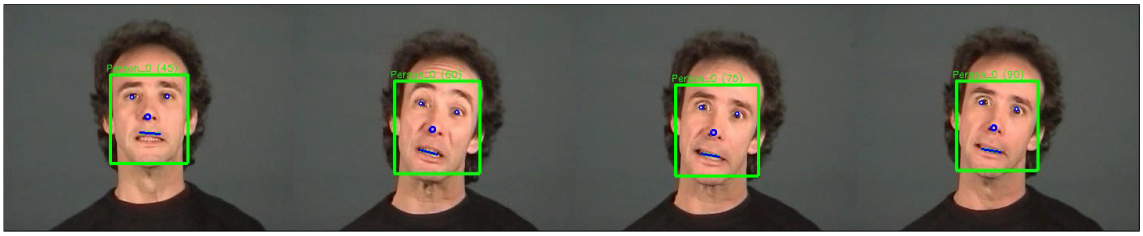


Figure 4.9: Face and facial element detection results for some samples of a sequence extracted from DaFEx [11].

As discussed in Sec. 4.4.4, an AAM realizes an iterative optimization scheme, which demands a reliable initialization. Neglecting such a suitable initialization might cause the AAM fitting process to fail completely. Such initializations are required when a system that is applicable in real-life situations is aimed at, the system should perform well in situations in which rigid head movements are encountered; as example. A bad alignment of the model derogate the fitting algorithm as well as the whole performance of the system [65].

To overcome this problem we proposed initialization methods that make AAM faster and more robust. The proposed initialization methods are based on the detected basic facial features; BFFs. Basically we use the mean shape $m = \begin{pmatrix} m_{x1} \cdots m_{xn} \\ m_{y1} \cdots m_{yn} \end{pmatrix}^T$ of the AAM as initial shape and place it within the detected face bounding box. The mean shape can be adopted to improve the fitting of the landmarks to the BFFs $f = \begin{pmatrix} f_{x1} \cdots f_{x4} \\ f_{y1} \cdots f_{y4} \end{pmatrix}^T$ (centers of right and left eye, nose and mouth). For each center of such a feature, there is a corresponding landmark in the mean shape. We refer to these special landmarks as *basic landmarks* $p = \begin{pmatrix} p_{x1} \cdots p_{x4} \\ p_{y1} \cdots p_{y4} \end{pmatrix}^T$, whereas all other points of the mean shape are simply called *landmarks*. Fig. 4.10 depicts the face bounding box as a white rectangle, the BFFs as white crosses, the basic landmarks colored green and all remaining landmarks in blue.

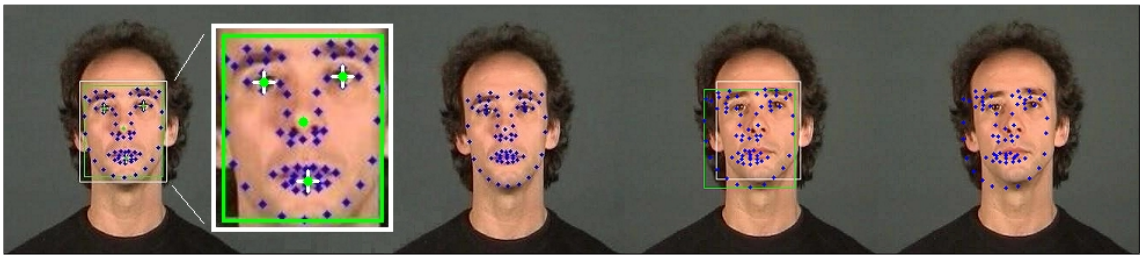


Figure 4.10: Initialization based on face bounding box and BFFs (first from left) and landmark matching via AAM search (second) for an image from DaFEx [11]. In cases where the initialization is too poor (third), the AAM search algorithm cannot eventually find a correct matching (fourth).

Since the detection component will not always robustly find all BFFs, the initialization works flexibly on any partial set given. If, for instance, only the bounding box (no BFFs at all) of the face is detected only a global scaling and positioning is applied. Given detected BFFs, the corresponding basic landmarks are adopted according to one of the following initialization schemes:

- **Linear transformation**: The size and position of the mean shape is linear transformed such that the distance between each BFF and the corresponding basic landmark is minimized:

  $m' = m \cdot \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} + \begin{pmatrix} d_x \cdots d_x \\ d_y \cdots d_y \end{pmatrix}^T$, where:
  $s_k = \frac{\sum_{i=1}^{4} \sum_{j=i+1}^{4} f_{ki}}{\sum_{i=1}^{4} \sum_{j=i+1}^{4} p_{ki}}$
  $d_k = \frac{1}{4} \sum_{i=1}^{4} f_{ki} - s_k \cdot \frac{1}{4} \sum_{i=1}^{4} p_{ki}$
  $k \in \{x, y\}$

- **Linear warping**: Each basic landmark is moved to fit the corresponding BFF exactly. The surrounding landmarks are also warped, depending on their distance to the BFF and the basic landmark. The displacement decreases linearly to the distance. Formally, for each landmark i, facial feature j and $k \in \{x, y\}$ do:

  $m'_{ki} = m_{ki} + d_{kij}$, where:
  $r = (m_{xi} - p_{xj})^2 + (m_{yi} - p_{yj})^2$
  $d_{kij} = (f_{kj} - p_{kj}) \cdot (1 - min\{\frac{\sqrt{r}}{w_k}, 1\})$
  w_k is a weight parameter

- **Gaussian warping**: Likewise to the linear warping, but the decrement of the displacement is Gaussian-based:

  $d_{kij} = (f_{kj} - p_{kj}) \cdot \exp(-\frac{r}{w_k})$

Sec. 6.3 will provide a comprehensive discussion about the relevance and appropriateness of each one of these initialization methods for the AAM fitting algorithm, and consequently for the performance of the whole system when facial-expression-based emotion analysis is aimed at.

4.7 Integration Concept in BIRON

The focus of our integration model is to enhance the performance of BIRON by providing the ability of inferring the facial expression displayed by its interlocutor for it. Toward this aim we adopt the face memory integration model and suggest concatenating it with an emotion understanding component. The face memory model was introduced by Hanheide et al. aiming to enable BIRON to discriminate between several persons living in a normal household environment, *"home-tour scenario"* [58]. Fig. 4.11 illustrates five basic components of the model. The first four components present the basic structure of the face memory system, while the fifth presents the suggested concatenation.

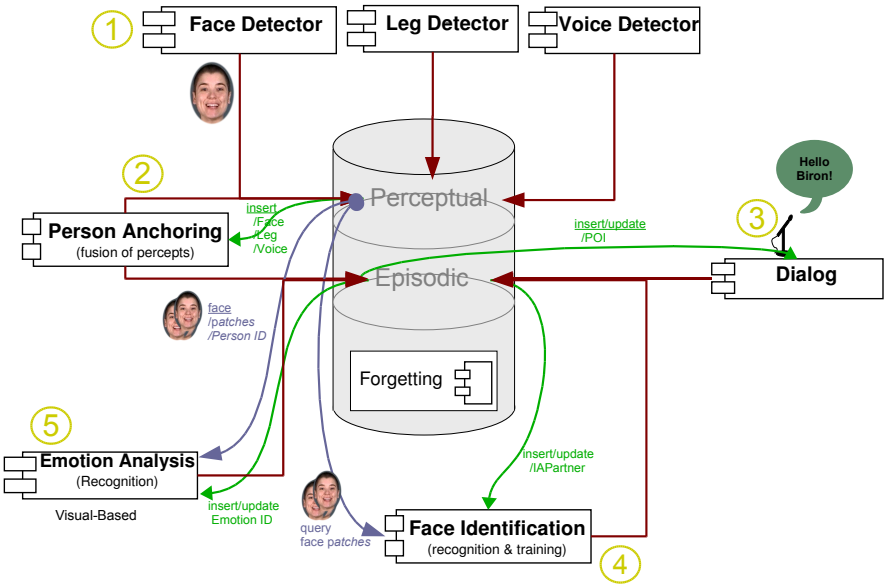


Figure 4.11: *Integration of facial-expression-based emotion analysis system in BIRON. Components 1-4 present the basic structure of the face memory model [58]. The fifth one presents the desired enhancement of providing an emotional ability..*

Considering the interaction between several components in the above-mentioned scenario, the first steps in our integration model are the same as in the face memory approach. These steps including the following: (I) a new POI hypothesis is submitted in the episodic memory as soon as the robot's sensors have detected the legs of the interactant, (II) the face detection component updates the POI in the episodic memory with face information and submits face views in the perceptual memory if the interactant's face is detected, and (III) the dialog component updates the POI with the information that the user is talking and submits information about the selected IP in episodic memory if an initiation phrase from the interactant is understood.

According to the name information provided by face recognition component, the facial expression component could update the IP with the emotional information in two separate ways. If the user is successfully classified the corresponding emotional profile is retrieved

from the memory to infer the current emotional state of the user (in our system one of seven predefined classes). Otherwise, if the identification of the user fails, a generic emotional model (including emotional profiles of all possible users) is retrieved to classify the expressed emotion. The dialog component can then retrieve the emotion-related information, which is updated by the facial expression analysis component, from the episodic memory and adapts the dialog method according to it.

Unlike the face recognition component, the facial expression classification one lacks online learning ability. Both person-dependent and independent emotional profiles should exist prior to the classification process. It is, however, feasible to get the facial expression classification process to carry out the face identification process in interactive learning of new unlabeled emotion classes, as long as the dialog component can obtain feedback about the encountered facial expression from the interaction partner himself.

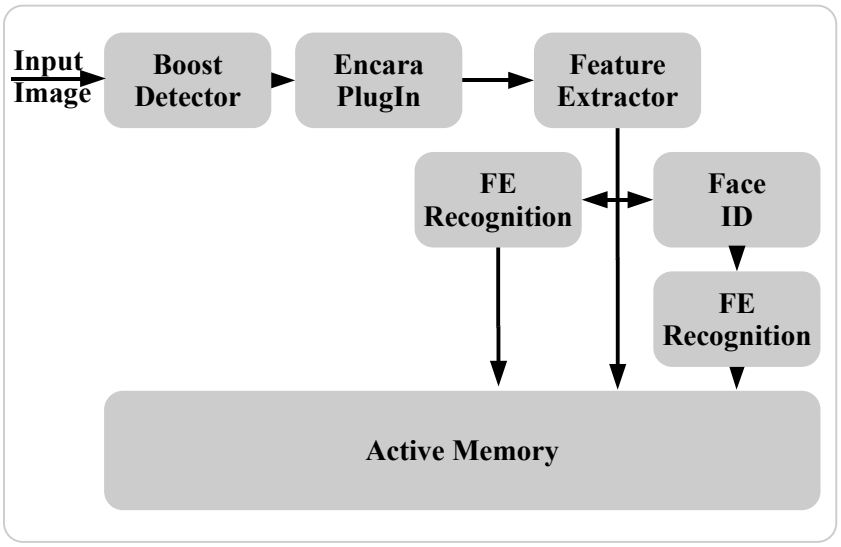


Figure 4.12: *Two work cases of a facial expression recognition system. FE: Facial Expression, ID Person Identity.*

Due to the realization that this integration model requires some architectural changes in BIRON's software, e.g., the current state of the dialog component lacks the ability to adapt according to the affective state of the user, we adopted a rather simple model, as depicted in Fig. 4.12. The model can work either completely independent from other components or according to the information about the identity of the user provided by the face recognition component. The input images are captured by the robot's camera and submitted in a primary face detector "Boost Detector" that outputs the image region (rectangle) in which the face is detected. Estimated face position is then used by Encara Plugin to detect some basic facial features, (eyes, nose, and mouth coordinations), which provide initialization for the facial features extractor. The extracted features are then submitted either into the facial expression classificator directly, or through the component of face identification. The components interact between each other through an active memory structure.

4.8 Summary

Facial expression may be the most convenient channel through which emotions can be displayed. However, facial expression analysis for inferring emotions is still challenged by several issues if the aim is to integrate this ability in a robot for social human-robot interaction in real-world conditions. The first challenge is that in real-world scenarios the interaction partner moves her/his head almost continuously. The interactant can move closer to or farther from the robot which affects the zoom and resolution of the detected face region, and furthermore the face of the user can be partially or even completely occluded. Another challenge is presented by the requirements that the system has to fulfill. The system should perform fully automatically from the first stage of detecting the user in the surrounding, to detecting his/her face, to extracting some facial features related to emotion displaying to deciding which emotion she/he is experiencing at the moment of capturing. Another requirement is that the system should be applicable in real-time conditions because any delay in detecting the emotional state of the user an reacting according to it will cause the system to be desynchronized and less efficient. Yet another challenge to systems for natural human-robot interaction applications is that they have to avoid any constrained conditions or any kind of manual preprocessing.

In this chapter we presented an integrated vision system for analyzing emotion conveyed through facial configuration. The developed system fulfills almost all the requirements of an ideal real-life visual-based emotion analysis system of being real-time applicable, fully automatic, and robust. For face acquiring, a fast and robust face detection approach is employed. This face detection scheme provides not only information about the position of the face in the captured image, but furthermore provides information about some facial features, eyes, nose and mouth, which can be used to align the facial extraction model.

A hybrid facial feature extraction approach, namely the Active appearance model (AAM), is used to extract some coherent features. AAM is concisely reviewed emphasizing its advantages by considering the variation in both shape and grey-scale of the face image in contrast to both geometric- and appearance-based methods, which on only one kind of variations.

AAM fitting method, however, can fail if rigid movements of the face are encountered. To overcome this problem an initialization scheme is proposed. Positions of the face and facial features, which are provided from the face detector, are used to define the initial location of an AAM. Different initialization methods are discussed in order to enhance the fitting speed and robustness of the AAM. An integration model of this system in our mobile robot as well a comprehensive evaluation on a sufficient database will be discussed in Sec. 4.7 and Sec. 6.3 respectively.

5 Audio-Visual Emotion Recognition

Multi-modal integration of affective information occurs during multi-sensory encoding and decoding. Humans articulate their emotions and perceive others' emotions through multiple modalities, such as speech tone, face configuration, and body movements. During multimodal communication these channels operate and interact dynamically all the time. Judgments for one modality may be influenced by other modalities either positively or negatively. One modality can provide further information about another or can increase their ambiguity.

A large body of psychological studies supported that facial expression and speech information are the most honest way to reflect the internal affective state on the outside [4, 125, 134]. A number of studies reported, however, the mutual influence between facial expressions and emotional tone of voice or emotional prosody [36]. Mehrabian stated that the semantic contents of a message contributes only 7% of the overall impression, while the major part of the information is embodied in nonverbal interaction, namely the vocal part and the facial expression contribute 38% and 55% respectively [94].

Furthermore, facial expressions have four further roles in addition to projecting the internal emotion. These roles are described by Wehrle and Kaiser [162] as:

- **a speech regulation signal (regulator):** the response of the listener tells the speaker that he can resume talking and if his words are understood or not
- **a speech-related signal (illustrator):** the speaker can raise his eyebrows in order to lay particular emphasis on his argumentation. The facial signals can also modify or even contradict the verbal messages, e.g., a smile can indicate that what is being said is not meant to be taken seriously
- **means for signaling relationship:** installing, maintaining, or aborting a relationship, e.g., when a couple is discussing a controversial topic, a smile can indicate that although they disagree on the topic there is no danger to the relationship.
- **an indicator for cognitive processes:** e.g., frowning often occurs when somebody does some hard thinking while concentrating on a problem, or when a difficulty is encountered in the task. And finally
- **an indicator for an emotion (affect display):** the person smiles because he is happy. Besides, affect that occurs during an interaction can refer to the interaction partner (becoming angry with others), but it can also refer to other persons or themes the interaction partners are talking about (sharing the anger about something)

Neglecting the multimodality aspect of affective communication will undoubtedly be very fallible so that, on one hand, the different modalities produce a significant amount

of complementary as well as redundant information that can be used to resolve problems when one of the modalities is not properly transmitted (e.g., speech in a noisy environment, or when the face of the interaction partner is occluded). On the other hand, dedicated emotions might be easier to read from one channel than from the others (e.g., sadness and fear are better identified through the audio channel, while anger and happiness are better identified through the visual channel [141].

Another thing is that relaying on only analyzing facial expression for emotion inferring in conversational sessions will be uncertain. That is because of the difficulty to distinguish between facial expressions that display emotions and those related to the speech processes. These processes can include lip movements, mouth configurations and the movements of the lower part of the face. Fig. 5.1 depicts the possible confusion in judging emotions, which are experienced during speech, with those displayed purely (without speaking). Furthermore, it is seldom that such situations are encountered, within which the information of both modalities is mutually independent (in conversational sessions, it is unnatural that the interaction partner speaks and displays emotions stepwise).

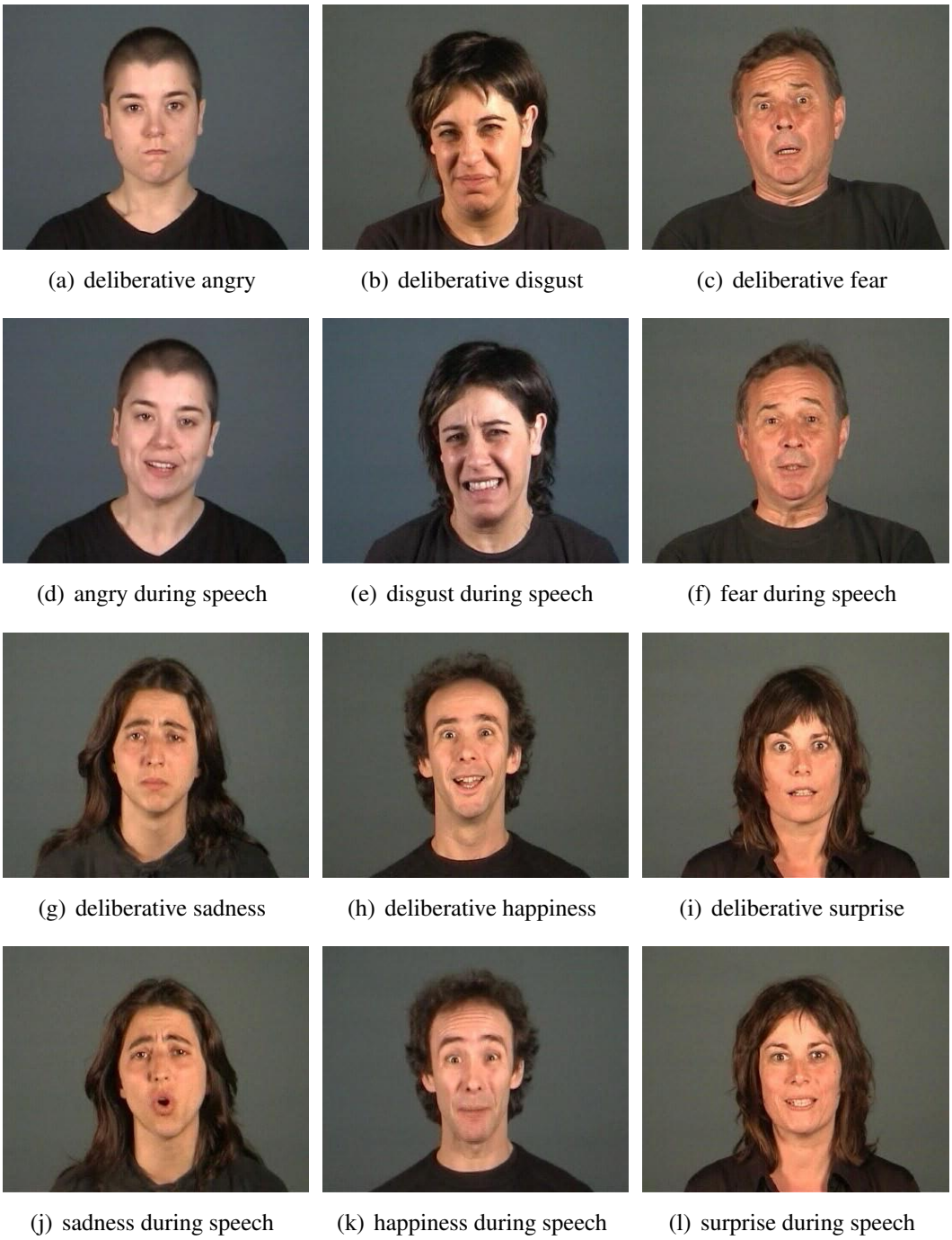
As in human-human interaction, multimodal recognition of emotion makes human-robot interaction more natural and efficient. As discussed in Sec 2.3, facial expression and speech information should play the major role in emotion expression. Hence, these two modalities should be considered as equal for multimodal human-robot interaction as they are for human-human communication.

While most current automatic emotion recognition approaches are uni-modal, in which the information processed by the computer system is limited to either face images/videos [85, 107, 117], speech signals [60, 157], or physiological measurements [16, 59], multimodal automatic recognition of emotions occurring in natural and real-life human-robot communication settings is still a largely unexplored and challenging problem.

Though, another requirement that has to be fulfilled by an ideal affective analysis system, in addition to those listed in Chapter 4, is to perform in a multimodal manner. Multimodality means that the system can handle two or more inputs, exp., facial expressions and speech tone, facial expressions and body movements, etc, simultaneously. Especially, joint analysis of facial expression and speech information should be addressed by designing such multimodal systems because:

(i) The recent findings of theoretical studies on emotion support the importance of the integration of information from different response components (such as facial and vocal expression) to yield a coherent judgment of emotions [91, 126]

(ii) To avoid possible realistic limitations of the current techniques of both computer vision and audio processing. For instance, current face analyzers are sensitive to head pose, occlusion, out-plan movements, and lighting changes while audio systems are sensitive to noise and distance between speakers and microphone. Moreover, if one channel fails for some reason, the other channel can still work. Thus, the final fusion performance can be more robust



(a) deliberative angry (b) deliberative disgust (c) deliberative fear

(d) angry during speech (e) disgust during speech (f) fear during speech

(g) deliberative sadness (h) deliberative happiness (i) deliberative surprise

(j) sadness during speech (k) happiness during speech (l) surprise during speech

Figure 5.1: The six basic emotions are conveyed by facial configurations in two conditions; deliberative (experience an emotion and displaying it without speaking), experience and displaying during speech session. The first and the third rows present the deliberative ones displayed by several individuals, while second and fourth rows present the same emotion of the same individual during speech. The judgment of the displayed emotion is confused in rows two and four if only the facial expressions are considered. Images are extracted from DaFEx database [11].

(iii) Combining facial expression information with acoustic signal cues helps to smooth the effects of speech-related facial configurations on facial expressions that display

emotions

The goal of this chapter is to explore new ways of human-robot interaction by enabling the robot to be more aware of the human user's emotional expressions. In particular, we concentrate on the problem of integrating audio-visual inputs for the detection the users' affective state conveyed by his/her facial expression and speech information at the same time. The aim of this combination is to smooth somewhat the effect of speech-related facial configurations on displaying emotions using facial expressions.

Facing the fact that facial configurations are influenced by both internal affective state and speech content, we apply a method that benefits from combining both complementary and conflicting information of both modalities. In this chapter we challenge this approach by analyzing the auditory and visual stimuli with respect to their general discriminative power in recognizing emotions. In the fusion stage, a probabilistic-based method is applied to combine audio and visual modalities so that the final affect recognition accuracy is notably improved.

This chapter first introduces briefly the relevance and related work on audio-visual emotion recognition in Sec. 5.1. Small introduction to emotion recognition from speech and an illustrative description of a speech-based emotion recognition system, which is used to infer emotions conveyed by acoustic modality, will be presented in Sec. 5.2. Three possible fusion methods of multi sensory data, which are dominant in this field, will be discussed in section 5.3. Afterwards, a probabilistic audio-visual data fusion approach is proposed in Sec 5.4, which is proceeded by a brief introduction to Bayes nets that constitutes the base of it. Sec. 5.5 will present a simple integration concept of an audio-visual emotion analysis system in BIRON. A comprehensive study using the DaFEx database, which evaluates the performance of the proposed multimodal system on recognizing the six basic Ekmanian emotions (anger, disgust, happiness, fear, sadness, and surprise) plus a neutral class, will be presented in Sec. 6.5. The ability of employing the bimodal system in real-life human-robot interaction will be discussed through an evaluation in life-like conditions, Sec. 6.6.

5.1 Related Work

Emotion analysis, using multimodal information, has been the subject of great deal of research in recent years. Paleari and Lisetti proposed a general framework for multimodal information fusion towards multimodal emotion recognition. They discussed that the fusion of the information takes place at signal, feature and, decision levels. However, the work did not report any practical implementation and experimental results [106]. Two integration methods, namely decision-level and feature-level, are discussed in the work of Busso et al. for fusing together facial expression and audio data [20]. For facial expression analysis, spatial data from predefined markers in each frame of the video is considered in order to extract a 4-dimensional feature vector per sentence, which is then used as input to the classifier. While for audio data, the means, standard deviations, ranges, maximum values, minimum values and medians of the pitch, energy, and

voiced-speech/unvoiced-speech ratio are computed. De Silva and Chi exploited a rule based method for decision level fusion of speech and vision based systems. The multimodal results showed an improvement over both of the individual systems [140]. Zeng et al. used a voting method to combine output of audio-based and vision-based recognition systems for person-dependent emotion recognition [172].

Castellano et al. used face, body and speech features. Speech has the highest rate of unimodal recognition accuracy. This may be due to the fact that the authors used pseudo-linguistic fabricated sentences as speech phrases and subjects said the same sentence with all emotions. The multimodal recognition accuracy is considerably higher than for any of the unimodal systems [21]. Mansoorizadeh and Charkari compared feature-level and decision-level fusion of speech and face information. Although both approaches had higher accuracies compared to the unimodal systems, the decision-level fusion showed to be more efficient than the feature-level fusion [90].

In a study by Song et al. three signals perceived from the subject, namely speech, facial expression and visual speech signals, are combined. The Facial Animation Parameters (FAPs) compliant facial feature tracking based on GASM (GPU-based Active Shape Model) is performed on the video to generate two vector streams which represent the expression feature and the visual speech feature. To extract effective speech features, they embedded the high-dimensional acoustic features into low dimensional space, which are then combined with the visual vectors in terms of a low-dimensional feature vector. A tripled Hidden Markov Model is then employed to perform the recognition. For facial feature tracking, however, all images in the used database should be labeled with some fiducial points [143].

5.2 Emotion Recognition from Speech

Since antiquity, people have realized the importance of vocal cues in the expression of emotion, as well the powerful effects of vocal emotion expression on interpersonal interaction and social influence. Speech is one of the indispensable means for sharing ideas, observations, and feelings. Automatic recognition of emotion based on a speech signal is an intensively studied research topic in the domains of affective computing [64, 122, 156, 157]. Recognition of emotional state is an increasingly important area in automated speech analysis due to several potential benefits that result from correct identification of subject's emotional state. As our focus in this work is natural human-robot interaction, correct assessment of interactant's emotion will improve the efficiency and the friendliness of human-robot interface.

Emotions, however, can be conveyed by speech information either explicitly through linguistic messages, (emotion-related words, happy or angry) or implicitly through (para-linguistic) messages that reflect the way the words are spoken. Some information about the speaker's affective state can be inferred directly from the surface features of words, which are summarized in some affective word dictionaries and lexical affinity (e.g.,

Whissell dictionary of affective language)[1], while The rest of the affective information lies below the text surface and can only be detected when the semantic context (e.g., discourse information) is taken into account.

Attending to only the verbal part (linguistic message), without regarding the manner in which it was spoken (paralinguistic message), will lead to some extent of loss of important aspects of the pertinent utterance and even to misunderstanding the spoken message. Furthermore, anticipating a person's word choice and the associated intent is very difficult, even in highly constrained situations as different people choose different words to express exactly the same thing.

When it comes to implicit, paralinguistic messages that convey affective information, the research in psychology and psycholinguistics provides a large body of results on acoustic and prosodic features which can be used to encode affective states of a speaker [32, 133]. However, researchers have not yet identified an optimal set of voice cues that reliably discriminate among emotions.

The general process of emotion recognition from speech signals can be sketched as following: (I)- speech signal capturing and preprocessing, (II)- extracting some emotion-related acoustic features, (III)- reducing feature dimensionality to an appropriate range suitable for classificator[2] and, (IV)- recognizing emotions with a suitable classifier. Speech signal preprocessing can include end-point detecting, separating voiced from unvoiced units, dividing signal into frames with predefined length, etc.

For an automatic recognition task, the signal to be recognized should be first characterized by measurable parameters, normally called feature extraction. The aim of feature extraction is to select some emotion-related features, which can set the base of a good classification. In the field of emotion recognition from speech, a variety of acoustic features are explored. For example, Banse et al. examined acoustic profiles or vocal cues for emotion expression using actors' voices for fourteen emotion categories. The exploited acoustic parameters were related to fundamental frequency/pitch (F0), energy, speech rate, and spectral information in voiced and unvoiced portions [4].

However, the features extracted from an audio signal can be categorized into two basic families: (I) basic features, which can be directly extracted from the signal itself, and (II) indirect features which can be extracted after applying some mathematical transformations on the original audio signal [122].

The family of basic features, in turn, can be divided into three further classes: features related to pitch, energy, and temporal behavior. The features of the first subclass can be presented by some values of audio signal pitch, such as the values of mean value, maximum, minimum, median, standard deviation, range (difference between maximum and minimum), variance value and change rate. Signal-energy-related features are usually used to present the power of the audio signal. These features can include values of mean, minimum, maximum, median, variance, range, and standard deviation. The third subclass of includes features that capture the temporal characteristics of the considered audio

[1] http://ketch.usc.edu/ abe/emotion_in_text_cgi/DAL_app/

[2] reduction of feature dimensionality is sometimes required when the computational complexity is considered

signal, such as cross zero rate features, segment length, and speaking rate [122, 157].

A spectral representation of a speech signal; Mel-scale Frequency Cepstral Coefficients, (MFCC), is suggested as a crucial feature for emotion recognition in real-life conditions [156]. MFCC is a parametric representation of an audio signal that is commonly used in the applications of automatic recognition of emotions. In order to calculate MFCCs a discrete Fourier transform is applied on each windowed segment of the signal, the power of the spectrum obtained from the previous step is mapped onto the so-called Mel-scale using N triangular-shaped filters. The powers of each Mel frequency are then logarithmised. A discrete cosine transform is then applied on the logarithmised Mel powers as if they were signals. Finally MFCCs are presented by the amplitudes of the resulting spectrum of the last step. Fig. 5.2 depicts some examples on the variation of audio signals and the extracted MFCCs when several emotions are uttered.

In addition to the requirement of being able to select and extract some features that present the encountered emotion the best, a reliable acoustic-based emotion analysis system should perform fully automatically. The segmentation of the incoming acoustic signal into meaningful units should be taken in account when such a reliable system is aimed for. As on one hand, emotion changes can occur very quickly, but the segment length sets the temporal resolution of recognizable changes, and, on the other hand, reliable statistical features can often only be computed over longer segments, the used system should find a reliable trade-off between these two issues [156]. Nevertheless in the evaluation study discussed in Sec.6.4 the whole utterance is considered as a unit to be classified.

To build our bimodal emotion analysis system, we exploited a framework called *"Emovoice"* and depicted in Fig. 5.3. The framework is introduced by Vogt et.al, to analysis emotions expressed through the acoustic channel. To segment the speech signal in life-like scenarios Emovoice exploited an algorithm called *"voice activity detection"*, which segments the signal into signal chunks of voice activity by considering pauses no shorter than 200 ms, shorter pauses will be omitted. This method is very fast and comes close to a segmentation into phrases though it does not use any linguistic knowledge [157].

Basic measurements extracted by EmoVoice are logarithmised pitch, signal energy, Mel-frequency cepstral coefficients (MFCCs, 12 coefficients), the short-term frequency spectrum, and the harmonics-to-noise ratio (HNR). The resulting series of values are transformed to different views, and for each of the resulting series mean, maximum, minimum, range, variance, median, first quartile, third quartile and interquartile range are derived [157].

The transformations into different views comprise the following:

- **logarithmised pitch:** the series of the local maxima, local minima, the difference, slope, distance between local extrema, the first and second derivation, and of course the basic series.

- **energy:** the basic series and the series of the local maxima, local minima, the difference, slope, distance between local extrema, first and second derivation as well as the series of their local maxima and local minima.

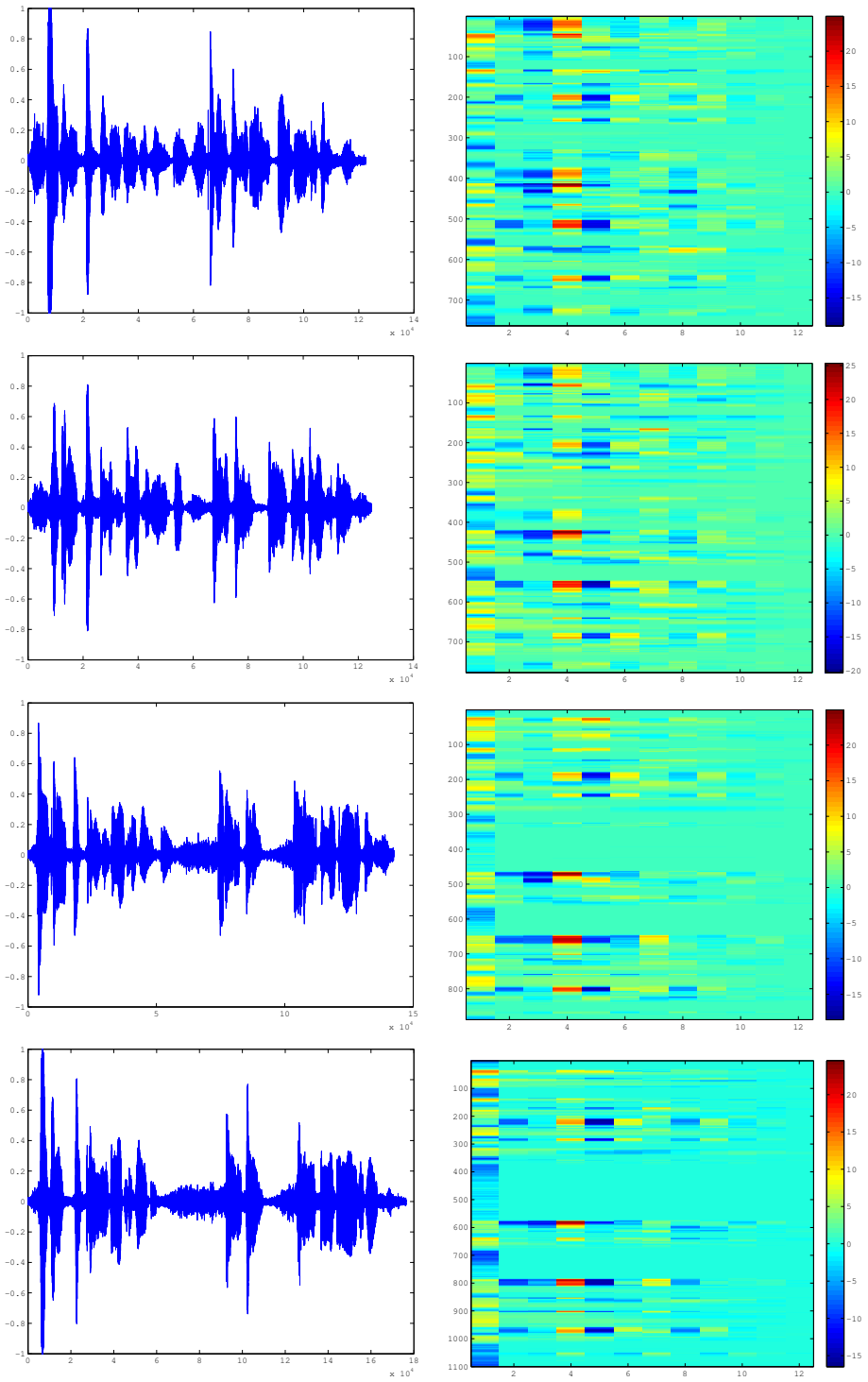


Figure 5.2: Emotion related variations in both original audio signals (left column), and the corresponding 12 MFCCs (right column). The original signals are presented in the time domain; x-axis presents duration in msec, and y-axis presents the amplitude. Variations related to anger, happiness, neutral, and sadness are presented here from top to bottom. The sentence pronounced in each utterance was in Italian *"In quella piccola stanza vuota c'era però soltanto una sveglia"*; in that little empty room there was only an alarm clock. Dafex database [11]

- **MFCCs:** the basic, local maxima, local minima for basic, first and second derivation for each of 12 coefficients alone
- **frequency spectrum:** the series of the center of gravity, the distance between the 10% and 90% frequency quantile, the slope between the strongest and the weakest frequency, the linear regression.
- **Harmonics-to-noise ratio, HNR:** only the basic series.

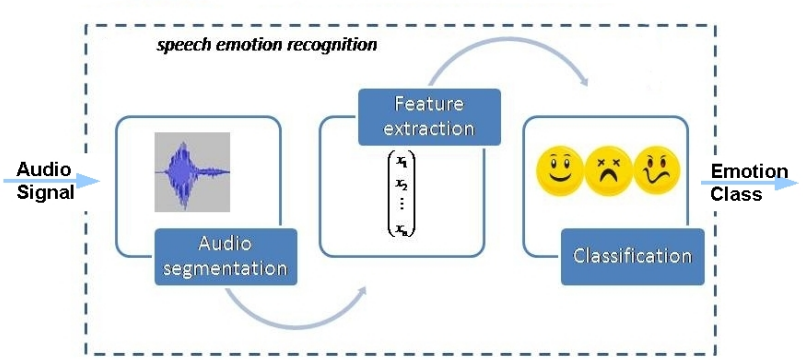


Figure 5.3: *Schematic Architecture of Acoustic-based Emotion Analysis System.* First stage, left, separates voiced units from unvoiced ones (pauses). The middle stage extracts some emotion related features from voiced units. Features are assigned with an emotional label in the third stage, furthest right. Printed by courtesy of Vogt [157]

To the features extracted above, segments length in seconds, proportion of pause to segment length, the number of the voiceless frames, and the speaking rate (distance between the global maximum and minimum of the segment) are added from the duration-related features. Additionally, pitch related features, such as positions of the global maximum and minimum in the segment and the number of local maxima and minima are considered too. From energy-related features, position of global maximum and the number of local maxima are inserted. Furthermore, jitter, shimmer and the number of glottal pulses of the analyzed speech features are taken in account as voice quality features.

From two classification methods that have already been implemented in Emovoice, namely a naive Bayes classifier (NB) and a support vector machine classifier (SVM), we have selected the latter for our bimodal system, which showed to perform equally well as the former when real-time applicability is considered, but outperform it when it comes to the accuracy of classification [157].

5.3 Fusion of Multisensory Data for Emotion Recognition

Multimodal information fusion is the task of combining some interrelated information from multiple modalities. In an emotion analysis system, while a unimodal system in-

corporates features of a single modality (visual, audio, tactile, or body information) the multimodal systems use information from multiple different modalities simultaneously.

However, theories of modality fusion in human perception do not agree on how information from different modalities should be integrated. For example, the Fuzzy Logical Model of Perception (FLMP) [91] stated that stimuli from different modalities should be treated as independent sources of information and be combined regardless of the kind of information they contain. This view is not undisputed (i.e. [36]) and it has been argued that the FLMP does not work well when confronted with conflicting information from different modalities [136]. Perceptual results suggest that, at least for the case of emotion recognition, the modalities should be weighted according to which information they convey best [46]: the visual modality primarily transmits valence (positive or negative value), whereas the auditory channel mainly contains information about activation.

In current fusion research, three types of multi-modal fusion strategies are usually applied, namely data-/signal-level fusion, feature-level fusion, and decision-level fusion. Fig 5.4 depicts the three possible levels of multimodal information fusion. Signal-level fusion is applicable solely to sources of the same nature and tightly synchronous. Generally it is achieved by mixing two or more physical signals of the same nature (two auditive signals, two visual signals of two cams, etc). This type of mixing is not feasible for multimodal fusion due to the fact that different modalities always have different captors and different signal characteristics (auditive and visual).

Feature-level fusion means concatenation of the features outputted from different signal processors together to construct a joint feature vector, which is then conveyed to the affect analyzer. It is used when there is evidence of class-dependent correlation between the features of multiple sources. For example, features can be extracted from a video processor (facial expression) and speech signal (emotion-related prosodic features). Feature-level fusion benefits of interdependence and correlation of the affective features in both modalities but is criticized for ignoring the differences in temporal structure, scale and metrics. Although, feature-level fusion demands synchronization of some extent between modalities. Another drawback of such a fusion strategy is that it is more difficult and computationally more intense than combining at the decision level. This is because of the increasing feature vector dimension, which consequently influences the performance of the whole system negatively [106, 171].

The third fusion strategy combines the semantic information captured from the individual unimodal systems, rather than mixing together features or signals. Due to the advantages of (I) being free of synchronization issues between modalities, (II) using relative simple fusion algorithms, and (III) their low computational requirement in contrast to the feature-based methods, decision-level fusion methods are adopted from the vast majority of researchers in the field of multimodality emotion recognition [90, 140]. Following this conclusion we decided a probablistic-based decision-level fusion method, which will be introduced bit later, to join the facial expression-based, and the acoustic information-based emotion recognizers into bimodal one [118].

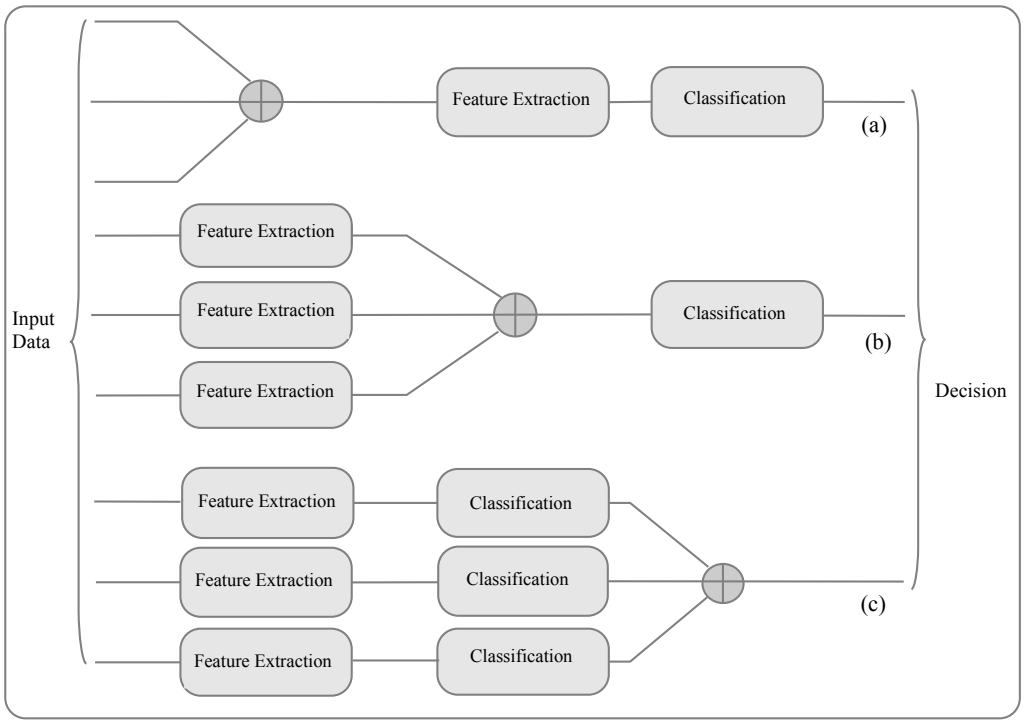


Figure 5.4: *Three basic fusion methods used in the current multimodal emotion recognition systems.* $\bigoplus$ is not a symbol of just simple adding. In (a) it means mixing two or more physical signals of the same nature. It presents basically the concatenating of features extracted from two or more sub-systems based on signals of different nature (b). And in (c) it can be any voting, weighting, or rule-based method.

5.4 Contribution

As seen above, the multisensory information can be fused in three different levels, namely at input, feature, and decision level. Due to the inherently different nature of our visual and acoustic cues, we decided on a decision-level fusion scheme. But instead of applying majority voting [172] or other simple fusion techniques, such as rule-based fusion by [140], we explicitly take the performance of each individual classifier into account and weight it according to its respective discrimination power. In the following, we introduce our fusion scheme preceded by a brief introduction to Baysian networks.

A Bayesian network is a graphical representation of the probabilistic relationships between a set of variables. Given a finite set $\mathcal{X} = \mathbf{X}_1, \ldots, \mathbf{X}_n$ of random variables, from which each variable $\mathbf{X}_i$ may take a value x_i from a specified domain, the corresponding Bayes nets consist of two basic components, namely the net structure and the local probability distributions associated with each variable. The network structure $\mathbf{S}$ is a directed acyclic graph (DAG) whose nodes correspond one-to-one to the random variables of $\mathcal{X}$. The second component describes the joint probability distribution of each variable in $\mathbf{S}$, given its parents.

In order for a Bayesian network to model a probability distribution, each variable

should be conditionally independent of all its non-descendants in the graph given the value of all its parents. That implies that the joint distribution of any variable in $\mathbf{S}$ can be decomposed into the following product form, by applying the chain rule of probabilities and properties of conditional independencies.

$$\mathbf{P}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \prod_{i=1}^{n} \mathbf{P}\big(\mathbf{X}_i | \mathbf{Pa}_i^S(\mathbf{X}_i)\big), \tag{5.1}$$

where $\mathbf{Pa}_i^S$ is the set of parents of $\mathbf{X}_i$ in $\mathbf{S}$.

As mentioned above, it is necessary to specify the parameters of the model, in addition to the graph structure. To achieve that, the conditional probability distribution (CPD) at each node of the directed net (DAG) should be specified. In the case of discrete variables, CPDs can be represented as conditional probability tables (CPTs), which list the probability that the child node takes on each of its different values for each combination of values of its parents. Consider the following example depicted in Fig. 5.5, in which all nodes are binary, i.e., have two possible values, denoted by T (true) and F (false):

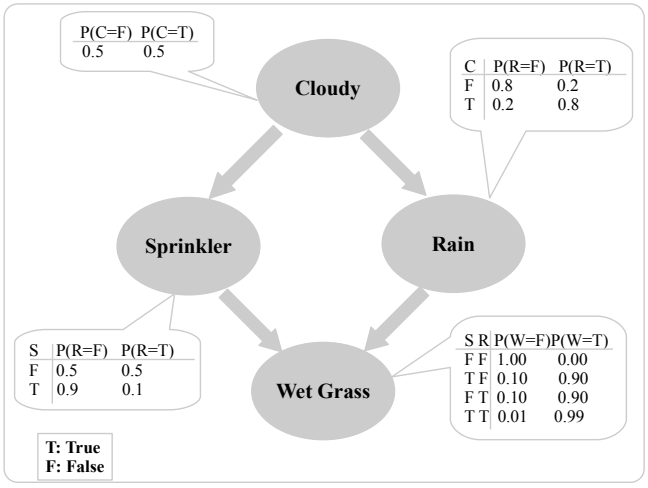


Figure 5.5: *An example of simple Bayes net with four random variables. The arcs encode the conditional dependencies between the variables. The example is derived from [128], P.p, 627.*

When the status of "grass" is observed, the event "grass is wet" (W=true) is caused by either the water sprinkler is on (S=true) or it is raining (R=true). The strength of this relationship is shown in the tables. For example, we see that $P(W = true|S = true, R = false) = 0.9$ (second row of the CPT of the node "Wet Grass"), and hence, $P(W = false|S = true, R = false) = 1 - 0.9 = 0.1$, because each row must sum to one. Since the C node has no parents, its CPT specifies the prior probability that it is cloudy or not (in this case, 0.5). The joint probability of all the nodes in the graph in Fig. 5.5 can be calculated by the chain rule of probability as following:

$$\mathbf{P}(C, S, R, W) = \mathbf{P}(C) \cdot \mathbf{P}(S|C) \cdot \mathbf{P}(R|C, S) \cdot \mathbf{P}(W|C, S, R) \tag{5.2}$$

And by using the conditional independence relationships, it can be rewritten as:

$$\mathbf{P}(C, S, R, W) = \mathbf{P}(C) \cdot \mathbf{P}(S|C) \cdot \mathbf{P}(R|C) \cdot \mathbf{P}(W|S, R) \tag{5.3}$$

This simplifying is allowed because R is independent of S given its parent C, and W is independent of C given its direct parents S and R.

The most common task that is solved by using Bayesian networks is the probabilistic inference. Suppose, for example, the fact that the grass is wet. This is caused by either it having rained, or the sprinkler having been on. Bayes' rule can be employed to compute the posterior probability of each explanation as the following (where $0 \equiv$ false and $1 \equiv$ true):

$$\mathbf{P}(S=1|W=1) = \frac{\mathbf{P}(S=1,W=1)}{\mathbf{P}(W=1)} = \frac{\sum_{c,r}\mathbf{P}(C=c,S=1,R=r,W=1)}{\mathbf{P}(W=1)} = \frac{0.2781}{0.6471} = 0.4297$$

$$\mathbf{P}(R=1|W=1) = \frac{\mathbf{P}(R=1,W=1)}{\mathbf{P}(W=1)} = \frac{\sum_{c,r}\mathbf{P}(C=c,S=s,R=1,W=1)}{\mathbf{P}(W=1)} = \frac{0.4581}{0.6471} = 0.7079$$

The term $\mathbf{P}(W=1) = \sum_{c,r,s}\mathbf{P}(C=c,S=s,R=r,W=1)$ is a normalizing constant that presents the probability (likelihood) of the data.

To fuse both individual modalities in a bimodal one we proposed a probablistic approach based on a top-down-reasoning Bayesian network with a rather simple structure depicted in Fig. 5.6. Based on the classification results of the individual visual and acoustic classifiers, we feed these into the Baysian network as evidence of the observable nodes (Acoustic and Visual, respectively). By Bayesian inference the posteriori probabilities of the unobservable affective fusion (Fusion) node are computed as:

$\mathbf{P}(\text{Fusion} = e_f|\text{Visual} = e_v, \text{Acoustic} = e_a)$, where, e_f, e_v, e_a can belong to any one of seven emotion classes mentioned above, and taken as a final result.

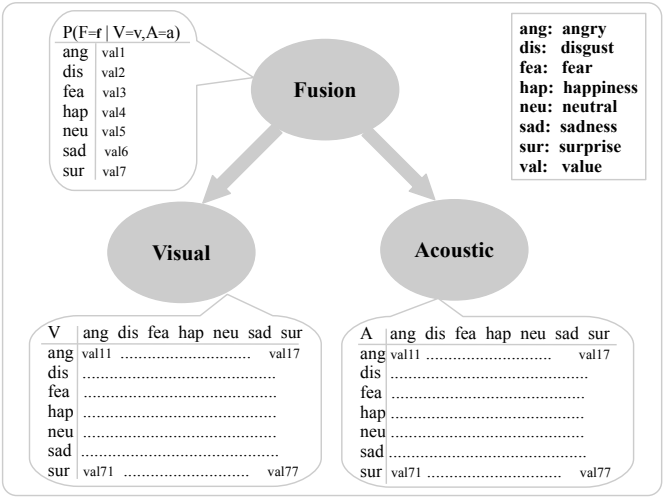


Figure 5.6: *The structure of the Baysian network used to fuse cues of both unimodals. Evidence of observable nodes – acoustic and visual – is fed as input into the corresponding node. The posteriori probabilities of the unobservable node are computed, with gives fusion as the final result.*

The required probability tables of the Bayesian network are obtained from a perform-

ance evaluation of each individual classifiers in an offline training phase based on ground-truth-annotated databases [12]. Therefore, confusion matrices of each classifier are turned into conditional probability tables modeling the dependent observation probabilities of the model according to the arrows in Fig. 5.6. In the notion of Zeng et al. [171], our fusion scheme is referred to as model-level instead of decision-level fusion, as it takes the respective classification performance models into account.

5.5 Integration Concept in BIRON

Challenging situations, in which the robots' interactant is engaged in a conversational course, we extended the model discussed in Sec. 4.7 in such a way that it combine the information provided by facial expression and that related to speech prosody. This combination is realized in order to smooth the negative effect of facial configuration related to the speech process on inferring emotions from facial expression. Fig. 5.7 illustrates a simple integration concept of audio-visual emotion analysis system in BIRON.

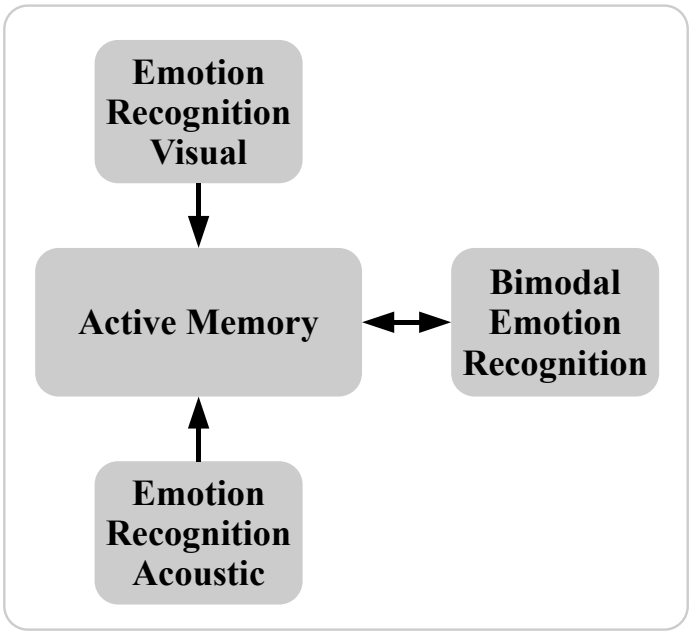


Figure 5.7: *Integration concept of both unimodals as a bimodal one in BIRON. Each system provides its own decision, which then fused together in the final decision .*

The speech-based system provides an utterance-based decision and inserts it in the memory. Accordingly, the facial-expression-based system is tuned to insert a majority voting decision for all video frames up to the last decision of the speech-based system into the memory. Using a decision-level fusion model, which is supposed to be already existing in the memory, both decisions can be fused yielding the final bimodal decision. More details about the fusion method can be found in Sec. 5.4.

5.6 Summary

In real-life conditions, situations are seldom encountered in which speaking and displaying facial expressions occur incrementally, i.e. the human speaks and experiences an emotion simultaneously rather than temporally separated. Challenging this point, the focus of this chapter was on how to draw benefits of several cues submitted by the speaker in such situations. More explicitly, how to combine facial-expression and speech-signal information in such way that provides the ability for the robot to recognize the interaction partner's emotion better.

As will be seen in Sec. 6.3, a degraded performance of a facial-expression-based analysis system is obtained, when an interaction partner is engaged in a conversational session. That is because of the ambiguity of distinguishing facial expressions and visual speech signals. In order to compensate this degradation, we fuse the prosody information extracted from speech with that extracted from facial configurations.

Three basic fusion methods are systematically analyzed: data-, feature-, and decision-level fusion. In data-level, the signals, which have to be mixed, should be of the same nature, either visual, or acoustic, and completely synchronized. The large size of the feature vectors in feature-level fusion demands relatively complex computation processes that consequently derogate the whole performance of the system, when the applications in real-life scenarios are aimed at.

As in our system facial-expression and speech-signal information are of an inherently different nature, a decision-level probablistic based fusion approach is proposed. In contrast to most current fusion methods of the decision level, our method explicitly take the performance of each individual modality into account and weights it according to its respective discrimination power.

The basic structure of the proposed methods is discussed in the last part of this chapter. The suitability of this method for combining facial expressions and speech signal cues in a bimodal emotion analysis system will be proven by a comprehensive evaluation in Sec. 6.5.

6 Evaluation and Discussion

This chapter provides a comprehensive evaluation of the performance of both the visual- and acoustic-based system as well as the bimodal one. As the focus of this work is to allow the robot to infer interactant's emotion in real-world conditions, i.e., the used technology is fully capable of online recognition of emotion either unimodal (facial-expression- or speech-information-based) or bimodal (audio-visual) system, an evaluation with real-life data is necessary. Obtaining such data, which is captured by the robot directly, is challenging due to several obstacles. On the one hand, facial expression analysis of any individual needs at least some facial images labeled with some fiducial points of this individual to have been included in a person-independent or in a person-dependent AAM. For unseen individuals, this inclusion process demands a tedious and time-consuming manual labeling of these images. Automatic alternatives present an open issue for future work. On the other hand, emotion recognition systems lack the ability of online learning, which is also considered as an area for further work. No online-learning ability means, that the emotional models (SVMs for visual- and acoustic-based systems, in addition to a validation matrix for the bimodal one) need to be already available in order for them to work in a reliable manner. In the case of employing our system in real-life application, without the above requirements being fulfilled, a degraded performance of the system is expected.

Thus, we present in this chapter an offline analysis of an actors database as previous work. A suitable dataset, that fulfils the requirements of being as natural as a real-life one and having a reliable ground truth, is used for this aim. The next two sections will provide an analytical discussion about current emotion databases, and their reliability and suitability to be used for evaluation. The performance of stand-alone facial-expression- and speech-information-based systems will be evaluated in Sec. 6.3 and Sec. 6.4 respectively, while the performance of the bimodal system will be evaluated in Sec. 6.5. As the focus of this work is to give the robot the ability of bimodal emotion recognition in life-like scenarios, the performance of all systems in real life conditions is evaluated in Sec. 6.6. A general discussion will conclude this chapter.

6.1 Emotional Databases

Designing reliable automatic affect recognizers requires adequate collection of labeled data of emotion expression for evaluation. Having such a sufficient database is challenged by two primary obstacles, namely data collection and labeling the collected data.

From the collection point of view, a large body of methods and strategies have been introduced for building such a database, according to the aim of the proposed system, the emotion relevant cues (facial expression, speech, body gestures, and physiological measurements), and many further structural aspects (technical equipments, environment

conditions, etc...). Some of the current databases, which are used to evaluate either stand-alone systems (vision-based and audio-visual) or bimodal ones (audio-visual), are listed in table 6.1.

Accurate description of what a person expresses and which feeling or affective state underlies this expression is considered to be the greatest obstacle to collecting such data. In general, three major approaches are used to judge the emotional states expressed via several cues (facial expression, acoustic signals): self-reporting, judging by external observer, labeling according to some specific changes in the face. While for describing the emotions displayed via facial expression all three approaches can be or even are already applied, see Table 6.1, only the first two approaches are valid for judging the acoustic-signals-related affective states.

6.1.1 Self-Report Approach

In self-report approach, the subjects are asked to report their feelings (usually retrospectively) and see whether their facial expressions or some emotion-related speech features differ when reporting changes in emotions. The emotions, however, are easy to experience but hard to explain *"Its meaning we know so long no one asks us to define it, Joseph LeDoux"*. Another drawback of such labeling approaches is of being error-prone, since subjects may fail to remember or distinguish among the emotions experienced, particularly if several minutes elapse before the report is made. *"A subject who successively felt anger, disgust, and contempt while watching a film might not recall all three reactions, their exact sequence, or their time of occurrence"* [44].

A few simplifications possibilities, however, have been proposed by researchers to avoid this problem. One of these simplifications is limiting self-report to the grosser distinction between pleasant (positive as happy) vs. unpleasant (negative as angry or sad) feelings or between deceptive and non-deceptive speech [62], but we then cannot determine whether facial expressions convey accurate information about particular unpleasant or pleasant feelings.

6.1.2 Judgment Approach

Judgment-based approaches are centered on the message conveyed by the considered cues (facial expressions, speech information). In order to categorizing affective state associated to one of these cues into a predefined number of emotion or mental activity classes, an agreement of a group of decoders is taken as ground truth. Each one of the two well-known emotion theories namely, basic emotion, dimensional emotion, has its own judgment approaches.

Most basic-emotion-based automatic emotion analysis systems attempt to cope with a database of facial images or sequences, audio recordings, or both which are directly labeled with one of a specific number of emotion classes [11, 87].

"Feeltrace"; a dimensional-emotion-theory-specific labeling tool, is a computer program implemented to let users describe perceived emotional content in terms of the two

well-known dimensions of "valence and arousal" [33]. The space is represented by a circle on a computer screen, alongside a window where a clip was presented. The vertical axis represented activation *"arousal"*; the horizontal axis evaluation *"valence"*. Raters used a mouse to move a cursor inside the circle, adjusting its position continuously to reflect the impression of emotion that they derived from the clip. The SAL database, listed in Table. 6.1 presents an example of a database that labeled using this tool.

6.1.3 Facial Configuration-based Approach

While the two above-mentioned labeling approaches aim to infer what underlies the displayed behavior, such as facial expression or speech signal, the sign-based judgment approach aims to describe the appearance, rather than the meaning, of the displayed behavior. There have been few attempts to establish objective coding systems which measure the positions of facial components involved in emotional expression and relate combinations of these measurements to the internal emotional states.

The most comprehensive and dominant sign-based encoding system in use is the *"Facial Action Coding System, FACS"*. FACS breaks each facial movement down into 44 action units *"AUs"* and attempts to describe the facial actions regarding to their location as well as their intensity [40]. Quite a few researchers have adopted this coding system to collect and label either image-based or video-based material to be employed for evaluating several facial expression analysis systems [9, 72, 111, 131].

6.1.4 Reliability of Labeling method

Indeed, the labeling of employed databases determines not only wether a given computing system attempts to analyze or interpret the emotion associated with specific signals recorded in the database, but may also influence the achievable recognition accuracy [149]. As seen above, an individual's emotional state can be judged either indirectly, by the observers' judgments (of the emotion experienced, the eliciting conditions, etc), or directly considering either the self-report or the measurement of facial activity (using any of the techniques described in the previous section).

However, when it comes to the real-time application, both self-report and facial-measurements labeling are not sufficient. This is because the former suffers under the timing issue and the latter from describing the shown facial changes while neglecting what underlies them, being relatively time consuming, and demanding professional-trained observers for labeling FACS-based data.

The most convenient method for labeling such reliable data is the judgment approaches. Indeed, that is not surprising because

(i) inferring basic emotions is an intuitive process and matches the experience of the ordinary human in daily life,

(ii) the simplicity of coping with a limited number of variables in contrast to other emotion models (e.g., 44 FACS),

(iii) their nature of being displayed and recognized universally, as proven in diversity of theoretical studies

(iv) avoiding the possible loss of information caused by labeling data into 2D or 3D space, and

(v) raters can be just ordinary individuals rather than the professional trained observers demanded for judging FACS.

6.2 Databases with Emotional Contents

According to the elicitation method of emotions, databases can be categorized into three major classes: induced, acted, and naturalistic. Naturalistic data seems the ideal way to collect data reliable for evaluating life-like affective systems, but the reality is not that straightforward. Having such data is challenging due to several aspects, such as problems of copyright and privacy, need of high developed tools to deal with it, and unreliable ground truth, are some obstacles challenging the employing of such a data to evaluate real-life emotion analysis systems.

Between the naturalistic facial expression, mentioned above, and the acted one, which will be discussed shortly later, lie various emotion induction techniques. There are various established methods such as listening to emotive music, looking at emotive pictures or films, and playing specially designed games. Such data, however, lack the certainty, as a specific induction methods *"emotive image"* could elicit disgust by some subjects, while it could trigger fear by others.

Most technological research on emotion continues to be based on recordings of actors, skilled [11] or not skilled [87, 131]. That is because of the difficulties of having naturalistic databases, on the one hand, and the unusability of induced facial expression databases to be employed for evaluating emotion analysis systems in life-like conditions, on the other. Therefore, we employed a dataset collected from skilled actors to evaluate the performance of our systems; there is a detailed description of the used data in Sec 6.2.2.

6.2.1 Emotional Databases in Use

In general, there is no comprehensive reference set of face images that could provide a basis for all of the different efforts in the research on automatic analysis of facial expressions. Only isolated pieces of such a facial database exist. Table 6.1 lists some already existing databases, which are used for evaluating the state-of-the-art visual- and audio-visual-based emotion analysis systems.

Quite a few of these databases contain solely static images. An example of such a database is the JAFFE database compiled by Lyons and Akamatsu. JAFFE contains in total 219 static images of 10 Japanese females displaying posed expressions of six basic emotions and is used for training and testing various existing methods for recognition of prototypic facial expressions of emotions [87].

Database Name	Type	Data Size	Emotion Description	Labeling	Human Evaluation
DaFEx [11]	AV	8 Subjects 336 Videos no Utterance 672 Videos with Utterance	6 Basic Emotions & Neural 3 Intensity Levels	80 Observers' Judgment	 76.8% 75%
Cohn-Kanade [72]	V	210 subjects X 3races 480 videos	6 Basic Emotions & AUs	FACS Observers' Judgment	-
AT&T [130]	V	40 Subjects 10 Images per Subject	Smiling Not Smiling	-	-
MMI [111]	V	61 Subjects 1250 Videos 600 Images	6 Basic Emotions Single AUs Combined AUs	FACS Observers' Judgment	-
BU-3DFE [1]	V	100 Subjects	6 Basic Emotions 4 Intensity Levels	-	-
Yale [54]	V	15 Subjects 11 images per subject	Sad, Sleepy and Surprise	-	-
Sebe et al. [138]	V	28 Subjects	Neutral, Happy Surprise, Disgust	Self Report	-
Fabo [56]	V	23 Subjects 210 Videos	6 Basic Emotion & neutral Uncertainty Anxiety and Boredom	-	-
JAFFE [87]	V	10 Japanese Models 213 images	6 Basic Emotions & Neural	60 Observers' Judgment	-
AR [2]	V	26 images per subject	Smile, Anger Scream and Neutral	-	-
CMU PIE [3]	V	68 Subjects 13 Poses 43 Illumination conditions	Neutral, Smile Blinking and Talking	-	-
CVL [4]	V	114 Subjects 7 Pictures per Subject	Neutral and Smile	-	-
Bosphorus [131]	V	150 Subjects 4666 Images	AUs Basic Emotions	FACS Observers' Judgment	-
RU-FACS [9]	AV	100 Subjects 100 Videos	AUs	FACS Two Coders	-
SAL [4]	AV	20 Subjects, one session 4 Subjects, two session	Dimensional Labeling	Feel-trace	-
CSC Curpos [62]	A	32 Subjects 3882 Speaking turns	Deceptive & Non-Deceptive Speech	Self Report	-

[1] http://www.lrv.fri.uni-lj.si/facedb.html
[2] http://cobweb.ecn.purdue.edu/ãleix/aleix_face_DB.html
[3] http://www.ri.cmu.edu/research_project_detail.html?project_id=418&menu_id=261
[4] http://www.lrv.fri.uni-lj.si/facedb.html
[5] http://emotion-research.net/toolbox/toolboxdatabase.2006-09-26.5667892524

Table 6.1: Overview of visual- and audio-visual-based databases of human affective behavior, V: only facial expressions are considered, AV: speech information and facial expression are taken into account, - means missing entry.

Image-based affective databases are important for obtaining information on the configuration of facial expression (which is essential in terms of emotions for inferring the related meaning). However, such databases are not sufficient for evaluating the performance of systems to be applied in real-world scenarios. Such systems demand motion records (video-based databases), which are necessary for studying temporal dynamics of facial expressions.

The MMI facial expression database was compiled by Pantic et al. [111]. It consists of two parts, namely part with deliberately displayed facial expressions and part with spon-

taneous facial displays. The first part contains over 4000 videos as well as over 600 static images depicting facial expressions of single AU activation, multiple AU activations, and six basic emotions. It has profile as well as frontal views, and is FACS coded by two certified coders. The second part of the MMI facial expression database currently contains 65 videos of spontaneous facial displays, that were coded in terms of displayed AUs and emotions by two certified coders.

The Cohn-Kanade facial expression database [72] may be the most widely used FACS-based database in research on automatic facial expression analysis [79, 85, 149, 152]. It is completely FACS orientated since not only the labeling is done according to the FACS system but also the subjects are instructed by the experimenter to perform specific single AUs or combinations of these. All desired displays of AUs are described and modeled prior to recording by the research scientist. The database consists of sequences of 9 to 60 frames, where each frame is stored as a single image. Sequences start at neutral and change gradually until they reach the maximum intensity of the performed AU; from this point, they change gradually back until the neutral state is reached once again.

6.2.2 DaFEx Database

We used the Dafex database that was compiled using eight trained Italian actors [11] to train and test both stand-alone facial-expression- and speech-information-based systems as well as the audio-visual system. The DaFEx database consists of 1008 short video clips of eight Italian actors (4 male and 4 female). Each clip comprises a presentation of one of Ekman's six basic emotions plus the neutral one and lasts between 4 and 27 Sec. The DaFEx database is divided into six blocks, in two of which, namely block 3 and block 6 the actors present facial expression without speaking; in the remaining blocks the actors speak and display emotional behavior simultaneously. Each actor in each of these blocks performs the seven emotions three times with different intensities (high, medium, and low). Fig. 6.2.2 shows some facial expressions presented by several objects.

Considering the influence of speaking on the displayed facial expression (i.e the influence of speech processes on facial expression), as discussed in section 2.3.3, is an essential issue in collecting databases sufficient for training and testing affective systems that are employed in real-life human-computer interaction. DaFEx is one of few databases to have taken this point into account [9, 11].

In addition to that, DaFEx has already been evaluated by 80 human observers. That draws huge benefits compared to other databases as the performance of the human observer can be set as reference to be compared against the performance of our systems. To the best of our knowledge, DaFEx is the sole audio-visual-based message-based dataset that offers this ability, as depicted in Table 6.1.

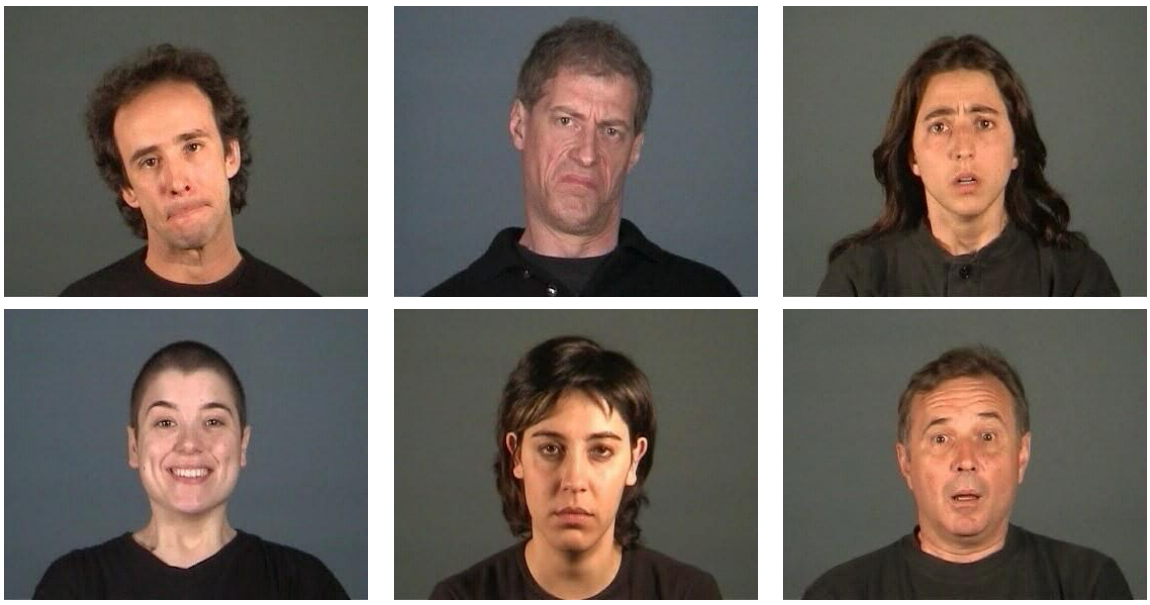


Figure 6.1: Six basic emotions presented by six different individuals; extracted from the DaFEx database [11]. The displayed emotions are, from left to right and top to bottom: angry, disgust, fear, happiness, sadness, and surprise

6.3 Evaluation of Facial-Expression-Based Emotion Analysis System

In order to evaluate the facial expression recognition system, the DaFEx database is used for training and testing. The third block of non-talking video data of each actor from DaFEx is selected to train and test the visual-based emotion analysis system in the case of solely displaying facial expressions. Some images of this block, with contains an average of 374.8 images from each actor (of each possible emotion class and each possible emotion intensity) are extracted and manually annotated. These images are then used to build a corresponding person-dependent active appearance model (AAM). The images from all actors are also used to build a generic,i.e., person-independent AAM. The parameter vectors of training data of each actor are then extracted twice firstly by using the corresponding AAM for each actor (person-dependent), and secondly by using the AAM that was built using data from all actors (person-independent). The extracted parameter vectors are subsequently conveyed to train a person-dependent and -independent support vector machine classifier respectively. SVM classifiers then categorize any unseen facial images into one of seven emotion classes (six Ekmanian plus the neutral one); the average amount of test data from each actor was 223.6 images.[1].

In both person-dependent and -independent cases a one-against-all SVM classifier with RBF kernel is trained. The following results highlight the robustness of our system as

[1] Prior evaluation of the facial-expression-based system can be seen in Sec. 8.1

	Ang	Dis	Fea	Hap	Neu	Sad	Sur	Total	Error
Ct	11.43	17.86	20.99	66.67	14.81	12.53	35.66	25.70	0.2124
BB	40.20	69.09	74.41	89.33	88.43	77.25	64.55	71.90	0.0489
LT	59.99	87.94	84.22	94.67	93.17	90.39	79.62	84.29	0.0540
LW	72.87	91.87	83.50	92.36	86.29	91.26	82.93	85.87	0.0485
GW	**75.49**	**90.20**	**90.96**	**94.20**	**92.25**	**93.97**	**83.83**	**88.70**	**0.0472**

Table 6.2: Recognizing rates obtained from the proposed facial-expression-based systems exploiting the person-dependent Active Appearance model. The system is evaluated on the DaFEx database. Emotions are; Ang:Angry, Dis:Disgust, Fea, Fear, Hap:Happiness, Neu:Neutral, Sad:Sadness, and Sur:Suprise. Initialization methods; Ct:Centering, BB:Bounding Box, LT:Linear Transformation, LW:Linear Warping, and GW:Gaussian Warping.

well as the impact of the initialization methods on the efficiency of the facial expression recognition subsystem; for more details about the proposed initialization methods recall Sec. 4.6.

As depicted in Table 6.2 the classification rates using the individual models, AAMs and SVMs, are higher than those when a generic one is used Table 6.3. That suggests putting forward the integration model in BIRON, which will be discussed in Sec 4.7. In this model, the facial expression component will benefit from the identity information provided from the face identification component. If the interaction partner is successfully identified, the corresponding emotional model, (SVM), can be used to recognize her/his emotions. Otherwise a generic emotional model can be utilized for unknown interaction partners.

The performance of the system, in the case of utilizing individual AAMs outperforms the performance with a generic one, because the variation of the facial features relevant to the expression of one individual is smaller than those of multi-person and the classes of individual models are clustered more compactly than the generic one. The column Error in both Table. 6.2 and Table. 6.3 indicates that the reconstruction errors of the former are larger than those resulting from the latter. That is, however, much expected because the larger the train data used for constructing an AAM, the smaller the reconstruction error [55].

When it comes to discussing the impact of the used initialization method on the performance of the system, the largest reconstruction errors and lowest recognition rates occurred when the model was aligned on about the image center (Centering). Coarsely initializing by using the bounding box already provided considerable enhancement of the performance. Minimizing the distance between the facial features and the feature points by using linear transformation initialization offered more adequate AAM fitting and therefore yielded better classification results. Moving the basic land-marks and their surroundings to fit the basic facial features, eyes, nose and mouth, according to either linear warping or Gaussian warping led to the best performance of the system [117].

The last rows of Table 6.2 and Table 6.3 depict the performance of the system in recognizing each individual emotion, given that a Gaussian-warping-based initialization method is used. From these rows, it can be seen that happiness achieved the highest recognition accuracy irrespective of being the used model, person-dependent or independent. That is however not surprising, when the judging rate of human observers is set as reference point [125]. Neutral achieved a relatively higher scores of 92.25% and 82.84% for the person-dependent and -independent model, respectively. Although neutral is not included as an emotional state in theoretical studies, deciding if the interaction partner is currently in an emotional state or not will definitely be beneficial for human-robot interaction. It could be a sign of the user engagement degree in the interaction course.

	Ang	Dis	Fea	Hap	Neu	Sad	Sur	Total	Error
Ct	22.70	10.95	06.55	99.50	05.56	07.34	24.09	25.31	0.1467
BB	38.63	39.95	68.46	99.33	84.37	71.43	67.44	67.09	0.0345
LT	81.23	70.55	71.74	98.61	78.91	78.41	73.79	79.04	0.0320
LW	77.38	63.76	76.61	99.33	83.60	82.76	84.73	81.17	0.0186
GW	**75.88**	**70.09**	**74.62**	**98.61**	**82.84**	**84.24**	**79.32**	**80.80**	**0.0180**

Table 6.3: Recognizing rates obtained from the proposed vision-based systems exploiting the person-independent Active Appearance model. The system is evaluated on the DaFEx database. Emotions are; Ang:Angry, Dis:Disgust, Fea, Fear, Hap:Happiness, Neu:Neutral, Sad:Sadness, and Sur:Suprise. Initialization methods; Ct:Centering, BB:Bounding Box, LT:Linear Transformation, LW:Linear Warping, and GW:Gaussian Warping.

To investigate the possible influence of facial configurations related to speech processes on the emotion-related facial expressions, we evaluated our visual-based system on the part of the DaFEx database that contains subjects speaking and displaying emotions simultaneously. Some annotated images of each actor displaying each possible facial expression with each possible emotion intensity from the first block of the DaFEx database; videos in utterance case, is selected to build a person-independent generic active appearance model (AAM) covering a total of 99% of the training set variance. The average was 244.66 images for each actor containing the expression of seven emotions in three different intensity levels. The parameter vectors for SVM training and testing are extracted from all four DaFEx blocks with utterance using this AAM. According to the outlined leave-one-out cross-validation scheme, training of one-against-all support vector machine classifiers with RBF kernel is conducted for person-independent classification of the visual appearance of the facial expressions into the six basic emotion classes in addition to the neutral one.

That the acoustic-based system in this evaluation was tuned to classify complete utterances, as will be seen in Sec. 6.4, a majority voting for each video sequence (mean sequence length: 80.47 frames) is applied. However, it shall be noted that for a real-time integration the problem of different segment sizes of the two cues needs to be ad-

dressed. While the facial-expression-based approach produces hypotheses for each frame, the speech-based approach needs a prior segmentation in order to determine a hypothesis.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Angry	**94.44**	02.78	00.00	00.00	01.39	01.39	00.00
Disgust	25.00	**73.61**	00.00	00.00	00.00	01.39	00.00
Fear	19.45	01.39	**58.33**	02.78	01.39	09.72	06.94
Happiness	15.28	00.00	00.00	**80.55**	00.00	01.39	02.78
Neutral	06.95	00.00	05.56	00.00	**79.16**	06.95	01.39
Sadness	11.11	00.00	06.94	01.39	05.56	**72.22**	02.78
Surprise	20.18	00.00	07.11	07.02	01.39	01.39	**62.91**
Total	**74.46**						

Table 6.4: Confusion matrix obtained from visual-based analysis of the DaFEx database in the case of talking subjects; rows represent the ground truth.

Table 6.4 depicts the confusion matrix of the classification results of the facial-expression-based system when the observed person speaks and displays facial expressions at the same time. The reason of unexpected confusion of disgust and surprise with angry might be that angry is not well presented by the actors and therefore the feature space of the angry utterances is spread widely. As in the case of displaying facial expression deliberately, neutral is recognized with a relatively high accuracy, which provides a significant enhancement towards our goal of better affective human-robot interaction. Speech mainly affects the mouth vertically more than horizontally; hence the horizontal deformation of the mouth can be related to the expression of emotions. For instance, happiness is partially expressed by horizontal extension of the mouth [41], and therefore it is still well recognized even when it is expressed during speech. During speech, most of the facial activities in the lower part of the face are related to lip movements, which considerably degrade the recognition rates of mouth-dependent classes, such as surprise. Surprise is associated with a wide open mouth [41], which makes it difficult to be distinguished from vertical mouth deformations that are associated with speech production processes. As downward motion of inner eyebrows and vertical wrinkles between them discriminate anger from other facial expressions [41], the high recognition accuracy of anger is highly expected. The reason of relatively high confusion of fear with surprise is that both emotion classes are characterized by the strong upward movement of the brows [41, 10]. A previous study suggested that fear and sadness are expressed in a similar movement in the forehead area [10], which could be the reason of the relatively high confusion betwen them; 9.72 point.

In order to compare the performance of the facial-expression-based system on talking and non-talking subjects, the same data of block one *"talking subjects"* as above is used to build a person-independent AAM. This AAM is then used to extract features of image data of block three and block six; block six contains video data with non-talking subjects

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Average
No Talking	97.22	69.44	86.11	97.22	80.56	69.44	94.44	**84.93**
Talking	94.44	73.61	58.33	80.55	79.16	72.22	62.91	**74.46**

Table 6.5: The results provided by the facial-expression-based analysis system in cases of the talking and not talking subjects. Majority voting scheme is used to calculate the classification rate for the whole utterance. As depicted the system has a better overall performance in the latter case than in the former.

too. The extracted features of each each block are then used to train the corresponding SVM model, which is then tested on the other block.

The classification results in cases of talking and no-talking subjects presented in Table. 6.5, indicate a degraded performance of the system when the observed person speaks and displays facial expressions simultaneously rather than displaying facial expression deliberately. The highest results were 84.93% for posed facial expression and 74.46% for facial expression and speech when the initialization method based on Gaussian warping is selected. The speech production process could be the major reason that leads to this degraded performance of the system. These process includes not only movements in the lower part of the face, such as lip movements and mouth configurations, but also some configurations of the upper part of the face, such eyelids and eyebrow movements.

6.4 Evaluation of Speech-Information-Based Emotion Analysis System

In order to smooth the effect of speaking on displaying facial expression, audio signals that are produced by speech processes are taken into account. To evaluate the speech-based recognition system, whole utterances from the same part of DaFEx are used as basic units of analysis, each utterance obtaining one emotion label. A leave-one-out cross-validation scheme is also utilized to train and test the speech-based system. The evaluation results indicate that the performance of the stand-alone speech system is significantly outperformed by the visual-based unimodal; 61.90% for the former and 74.46% for the latter [118]. This may be due to the DaFEx database being primarily designed for visual analysis and the speech recordings containing background noises, but it also emphasizes the need for a joint analysis.

The confusion matrix, depicted in Table. 6.6, shows an expected confusion of anger, happiness and fear, as all three are associated with high activation. However, it is noticeable that disgust, fear and surprise have a relatively high confusion rate with neutral. A previous study suggested that when neutral is excluded, anger and sadness are recognized by speech the best [133]. The results provided by our acoustic-based system verify this argument. This study suggests also that negative activation emotions, such as sadness, and the ones with positive activation, such as anger, happiness and fear, can be discriminated according to some values derived from the mean frequency,*"F0"*. The former set seems

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Angry	**68.05**	05.56	16.67	02.78	05.56	00.00	01.39
Disgust	05.56	**51.38**	06.95	15.28	06.94	12.50	01.39
Fear	15.20	08.33	**48.61**	12.50	00.00	09.72	05.56
Happiness	06.94	15.28	15.28	**50.00**	06.95	00.00	05.56
Neutral	00.00	06.95	01.39	01.39	**87.49**	02.78	00.00
Sadness	00.00	08.34	13.89	05.56	01.39	**69.44**	01.39
Surprise	04.17	06.94	13.89	11.11	02.78	02.78	**58.33**
Total	**61.90**						

Table 6.6: Confusion matrix obtained from speech-based unimodal analysis of the DaFEx database; rows represent the ground truth.

to be associated with decreased mean and range of mean frequency, while the latter is accompanied by increased mean and range of F0 [133].

6.5 Evaluation of Audio-Visual System

The individual results, illustrated in Table 6.4 and Table 6.6, indicate that the visual recognition system overall yields better results than the acoustic system. The total recognition rate of the former is 61.90% while of the latter it is 74.46%. This notion is verified by concordant evidence of several theoretical studies [133, 134]

However, when comparing the performance of each modality for each emotion, it can be seen that there are differences depending on the emotion as depicted by the ratios (Vis/Aco) between visual and acoustic performance in Table 6.7. In order to normalize against the overall performance difference, which we assume is due to artifacts of the data and probably also due to intrinsic properties of the modalities, we normalized these ratios by the overall Vis/Aco factor of 1.23, obtaining relative Vis/Aco ratios. Emotions with a relative ratio below 1 this indicates that the speech-information-based channel provides a relatively better performance, while those with a relative ratio above 1 indicate a higher performance of the visual channel.

The overall results provided by unimodal systems suggest that for accurate and reliable recognition of emotion classes the modalities should be combined in some manner that benefits the interrelationships between the individual classes and the underlying modalities. For a fusion scheme, it means that it would indeed be beneficial to integrate hypotheses from both modalities based on the a-priori confidence of a modality for a certain emotion and weight by the own confidence of the classifier in the current classification as proposed by the *"Fuzzy Logical Model of Perception, FLMP"* [91]. To do that, we put forward our fusion model, as proposed in Sec. 5.4

In order to evaluate our bimodal system, the same subset of DaFEx that contains only videos with speaking actors, namely (block 1, 2, 4, and 5), is chosen. Due to the small

	Vis	Aco	Vis/Aco	Relative Vis/Aco
Ang	94.44	68.05	1.35	1.10
Dis	73.61	51.38	1.43	1.16
Fea	58.33	48.61	1.20	0.98
Hap	80.55	50.00	1.61	1.31
Neu	79.16	87.49	0.90	0.73
Sad	72.22	69.44	1.04	0.85
Sur	62.91	58.33	1.08	0.88
Total	74.46	61.90	1.23	1.00

Table 6.7: Classification rates of the visual vs acoustic systems and their ratios.

sample size, the same actors are used for training and testing. It shall be noted, however, that both speech-based and facial-expression-based recognizers apply person-independent models. The same leave-one-out cross-validation is used for the different modalities. For each unimodal, training is done on three blocks and evaluation of the performance is performed on the one remaining test block. The probability tables for the Bayesian fusion model are obtained from validation of the performance on the three training blocks. The fusion performance is tested again on the test block. In cross-validation, all permutation of blocks are applied to training and testing respectively.

	Visual	Acoustic	Audio-Visual
Anger	94.44	68.05	81.94
Disgust	73.61	51.38	87.50
Fear	58.33	48.61	52.78
Happiness	80.55	50.00	86.11
Neutral	79.16	87.49	86.11
Sadness	72.22	69.44	74.99
Surprise	62.91	58.33	77.77
Total	**74.46**	**61.90**	**78.17**

Table 6.8: Classification rates of the visual, audio, and audio-visual systems

Table. 6.8 depicts the significant overall improvement achieved by the proposed fusion scheme of applying our simple Bayesian networks model, which has the advantage over the facial-expressions- and speech-information-based unimodal systems of about 4% and 16% points, respectively. The 2nd and the 7th rows (Disgust and Surprise, respectively) reveal a high accuracy of the fusion model for recognizing disgust and surprise, respectively in contrast to the stand-alone unimodals, indicating that both cues obviously com-

prise complementary information that facilitate eased discrimination in the joint analysis. On the other side, it is noticeable that both unimodal cues comprise only redundant information so that the fusion yields no improvement with regard to discrimination ability for the recognition of fear. Overall our system achieves good results on the DaFEx database, which are comparable with those reported for human observers [12], which indicates that the interpretation of facial expressions is a difficult task for humans, too.

	Ang	Dis	Fea	Hap	Neu	Sad	Sur
Ang	**81.94**	08.33	01.39	02.78	04.17	00.00	01.39
Dis	04.17	**87.50**	01.39	02.78	00.00	04.17	00.00
Fea	15.28	02.78	**52.78**	05.56	00.00	12.50	11.11
Hap	05.56	01.39	00.00	**86.11**	00.00	01.39	05.56
Neu	04.17	00.00	01.39	00.00	**86.11**	06.94	01.39
Sad	01.39	08.33	06.94	01.39	02.78	**74.99**	04.17
Sur	01.39	01.39	05.56	11.11	01.39	01.39	**77.77**
Tot	**78.17**						

Table 6.9: Confusion matrix obtained by using audio-visual system based on a probabilistic decision level fusion method; rows represent the ground truth.

The confusion matrix provided by the audio-visual system, as depicted in Table. 6.9, shows a high confusion of fear with anger and surprise, 15.28% and 11.11% respectively. Excluding the confusion values of fear, all other values do not exceed 10, which again emphasize the overall improvement of the audio-visual system compared with both standalone unimodals.

6.6 Evaluation in Real-Life Conditions

As we are striving in this work to give the robot a bimodal emotion recognition ability that is based on analyzing facial expressions and speech information, the same procedure of evaluation using the DaFEx database is repeated with subjects in a real-life conditions. Four subjects have participated in this test (one female and three males), some examples are presented in Fig. 6.2.

The whole procedure is divided into training and test phases. For one subject both phases were conducted in the same day; for two others the test was is conducted in the following day, while for the fourth subject the time interval was two days. In the training phase the subjects are asked to display facial expressions of five emotion classes: anger, happiness, neutral, sadness, and surprise. The average amount of data captured from each subject for each facial expression class was 246 images. To create conditions of real-life human-robot interaction as much as possible, the subjects are allowed to move arbitrarily in front of the camera. During this phase a person-independent AAM, which is built from

Figure 6.2: Examples of image data captured by robot's camera directly, anger and happiness are displayed in left and right image respectively

a subset of the DaFEx database of talking and non-talking subjects, is used to extract the emotion-related facial features. These features are then conveyed to train a person-dependent SVM. The test phase, in turn, is divided into two sessions. In the first one the subjects are asked to show the five facial expression classes as mentioned above, while in the second sessions they are asked to display facial expressions and utter a few sentences (in general five) expressing as much an emotions as possible [2]. In both sessions of the test phase, the above-mentioned AAM is used to extract facial features, which are labeled with the proper emotional class by the above-trained SVM. In the session of speaking and displaying facial expression a person-independent speech-based emotion recognizer is utilized to categorize each utterance into the proper emotional class. An average of 145.25 and 163.5 images from each subject for each emotion are used as test data in the sessions of talking and non-talking subjects, respectively. The validation matrix for the fusion scheme of each subject was an averaged confusion matrix (CPT), which is obtained from the performance of both individual systems on the three remaining subjects.

Table 6.10 illustrates the result obtained by using only the facial-expression-based emotion analysis system to recognize emotions that are deliberatively displayed by the subjects. As depicted in the table, the most negative emotion – sadness – and the most positive emotion – happiness – are recognized the best. Neutral also has a relatively high recognition rate, which can serve to distinguish between emotional and non-emotional states of the interactant. The mutual confusion between sadness and neutral indicates the similarity between them when the distinguishing is based only on analyzing the associated facial expressions. The fact that surprise is a transient state, difficult to hold, which changes rapidly into another one (in our test it changed generally into the neutral state), could be the reason for the relatively high confusion of surprise with neutral.

The results obtained by analyzing facial expressions during speech are illustrated in the table 6.11. The results present the recognition rates after applying majority voting for each

[2] The sentences were emotional words free

	Anger	Happiness	Neutral	Sadness	Surprise
Anger	**57.72**	00.60	12.19	28.54	00.95
Happiness	02.96	**67.46**	21.00	07.15	01.42
Neutral	05.21	00.00	**64.36**	30.42	00.00
Sadness	02.98	00.00	17.32	**79.18**	00.53
Surprise	05.57	00.88	31.35	10.55	**51.64**
Total	**64.07**				

Table 6.10: Confusion matrix obtained by using the facial-expression-based system in the test session of displaying emotions deliberatively; rows represent the ground truth.

	Anger	Happiness	Neutral	Sadness	Surprise
Anger	**75.00**	00.00	06.25	18.75	00.00
Happiness	25.00	**43.75**	25.00	06.25	00.00
Neutral	20.00	00.00	**50.00**	30.00	00.00
Sadness	22.36	11.11	06.25	**60.28**	00.00
Surprise	16.67	00.00	12.50	22.92	**47.92**
Total	**55.39**				

Table 6.11: Confusion matrix obtained by using the facial-expression-based system in the test session of expressing emotions via facial expressions and speech tone simultaneously; rows represent the ground truth.

utterance that doesn't include a pause longer than 200 ms. As in the evaluation with the database (offline evaluation), facial-expression-based analysis of emotion delivered lower recognition rates when the subjects were engaged in conversational sessions; 64.07% for recognizing facial expressions displayed deliberatively and 55.39% for recognizing facial expressions during speech. The higher recognition rate of anger during speech compared to anger displayed deliberatively could be because majority voting over the time of each sentence is applied in the former, while the recognition rate of the latter is computed for the entire video sequence.

Table 6.12 illustrates the results obtained from both the stand-alone and bimodal systems. The low rates delivered by the speech-based emotion analysis system - the second column - could be because a person-independent classifier is used, which is trained on a speech-based emotion database that does not include the subjects participating in the evaluation procedure. Nevertheless, it can be seen that the whole performance of the bimodal system has an advantage over both facial-expression- and the speech-information-based systems, which satisfy the goal of the fusion scheme proposed previously. However, when the performance of each channel on each emotion is considered it is notable that the recognition rate of happiness and neutral is enhanced when the bimodal system

	Aco	Vis	Vis/Aco	Relative Vis/Aco	Audio-Visual
Anger	33.04	75.00	2.27	0.805	75.00
Happiness	15.42	43.75	2.84	1.007	50.00
Neutral	36.25	50.00	1.38	0.489	68.75
Sadness	23.06	60.28	2.61	0.925	49.03
Surprise	10.42	47.92	4.60	1.631	47.92
Total	**23.63**	**55.39**	2.82	1.00	**58.14**

Table 6.12: The performance of each stand-alone unimodal systems, their relative performance on each emotion class, and the performance of the bimodal system. All results are obtained from a test in a real-life condition.

is employed, which indicates that the cues of both modalities comprise complementary information for these two emotions. In contrast, from the first and fifth rows, it is noticeable that both unimodal cues comprise only redundant information so that combining both modalities yields no improvement with regard to discrimination ability for the recognition of anger and surprise. Furthermore, the fourth row indicates that both modalities deliver conflicting information, which causes sadness to be recognized even less than the stand-alone facial-expression-based modality.

The comparison between the performance of all of the systems in the cases of offline (DaFEx database) and online (data captured in real-life conditions) evaluation shows better performance of the systems in the former case, especially of the speech-information-based system. These performance differences were greatly expected because (I) the speech-information-based system in the former was trained using data from the same subjects who had participated in the evaluation test, (II) the facial-expression-based system of the former case was trained and tested on a relatively constrained set of data (the actors displayed almost a frontal-view facial expression with constrained head movements while they were sitting in front of the camera), and (III) the degraded performance of both unimodal systems will consequentially lead to a degraded performance of the bimodal system.

6.7 General Discussion

Sofar, we have presented an audio-visual emotion recognition system to be integrated into a robot. But does this system relate to the developments in the field of social human-robot interaction? More precisely, how can understanding of interactant's emotion help for more social human-robot interaction.

Sec. 2.1 stated that emotions play a major role in human-human interaction. They seem to be centrally involved in determining most human's behavioral reactions to external and internal events of major significance for needs and goals of humans. For instance, Frijda suggests that positive emotions, exp., happiness, are elicited by events that satisfy some

motive, enhance one's power of survival, or demonstrate the successful exercise of one's capabilities. Positive emotions often signal that activity toward the goal can terminate, or that resources can be freed for other exploits. In contrast, many negative emotions, exp., anger and sadness, result from painful sensations or threatening situations. Negative emotions motivate actions to set things right or to prevent unpleasant things from occurring [52].

By mirroring this fact in human-robot interaction scenarios, BIRON can adopt according to the emotional state of its interaction partner, which might be previously elicited by a previous behavior of BIRON itself. As example, in *"object-teaching scenario"*[3], BIRON can detect a smile of its interactant as an acceptance sign, after it has recognized successfully a previously learned object, *"That is a book"*, and precedes capturing further features of the object, which can help BIRON in similar scenarios in future. In the so-called *"home-tour scenario"*[4], BIRON can infer that its interactant is getting annoyed in cases when it has not located itself correctly *"BIRON still insists it is standing in the kitchen, when in fact it is actually in the dining room"*, by detecting some features related to the anger emotional state from the interactant's facial expressions (frowning face), voice, or both. Accordingly, BIRON can try to fix such a situation by asking additional questions in order to ascertain whether it was right or, if not, what was wrong and how the problem can be solved. Detecting a surprise in the interaction partner could indicate that BIRON has performed an unexpected behavior, *"BIRON pronounces a phrase that is far removed from the current context of interaction, exp., you have a very nice living room, when an interaction partner is in fact displaying a bottle in an object-teaching scenario"*, which can be followed by a question from BIRON to determine the reason behind the interactant looking surprised, and to work out how this can be avoided in future. Getting a neutral state of the interaction partner for a specific time period could indicate an interaction with a low level of engagement, BIRON could suggest having a cup of coffee or playing a game.

As discussed in the previous section, the proposed audio-visual emotion analysis system provides the ability for BIRON to infer all emotional states with a relatively high accuracy. That allows BIRON to perform in interaction scenarios similar to those discussed shortly before.

[3] more details can be found in Sec. 8.3
[4] more details can be found in Sec. 8.2

7 Conclusion and Future Work

Modern robots are developed in such a way that they won't just function inside factories but can also take part in our daily life. They could work in normal everyday life environments and interact with even non-expert users. To maintain a natural human-robot interaction, the robot needs to understand the human via different observational modalities. In this dissertation we focus on the audio-visual recognition of human emotions. Emotions are associated with several internal changes inside the human. Projecting these changes onto the outside via various media, such as for instance facial expression and speech prosody, as well as sensing emotion-related cues submitted by others are often considered as the main part of social interaction. Acting according to this fact, it is suggested that sensing the emotions of the interaction partner is essential for social human-robot interaction, and that inferring the emotions by the robot is obvious for efficient and user-friendly human-computer interaction. This inference ability is especially needed when more sophisticated emotional behavior of a robot towards its interactant is intended, e.g., behavior adaption of the robot according to the emotional state of its interaction partner.

However, most current techniques for the recognition of emotions cannot be transferred because they rely on only one modality to infer the encountered emotion neglecting that humans encode their own and perceive others' emotion multimodally. Furthermore, quite a few of these approaches lack the abilities to be applied fully automatic in natural and human-human-like social human-robot interaction.

The presented work overcomes several of these deficiencies. In order to infer six basic emotions of the interactant plus the neutral one, we presented an integrated vision system based on analyzing the associated facial expression. For realizing this system a hybrid facial features extraction method is employed. To provide the ability for the robot to recognize the emotion fully automatically in a natural and unconstrained environment, a novel initialization method is proposed.

This initialization method aimed at benefitting from basic facial features supplied by the face detector to initialize the location of the feature extraction model. Because of the lack of a real-life data with reliable ground truth that captured by the robot directly, the system is trained and tested on a sufficient, and ground-truth annotated database. The results presented that the information related to these basic features as well as the way in which it is used have a great impact on the performance of the feature extractor and consequently on the performance of the whole system. The results evidenced, furthermore, that facial expressions are better classified, when a person-dependent model is utilized than when using a person-independent one. In order to make use of this point, the system has been developed in such a way that facial expression benefits from the prior step of user identification.

In natural human-robot interaction, however, reliance is not only placed on visual ob-

servations for sensing and recognizing the emotional state of the user, but rather it occurs multimodally in the most natural way via facial expressions and speech cues. Most current emotion sensitive systems employ either the former or the latter. The few approaches that focus on multimodal emotion recognition employ several cues in rather simple ways neglecting the mutual influence between them. Facing this challenge, we presented an audio-visual-based emotion-aware system that focuses on the emotion analysis of talking interlocutors, which is different to most approaches which focus on non-talking faces-Like the stand-alone visual-based system, the bimodal system fulfils the requirement of being fully automatic and having real-life applicability. A probabilistic-based decision-level fusion approach is introduced to combine the cues of both unimodals. Being based on Bayes nets, the used fusion method draws benefits by taking the performance of each individual classifier into account and weighting them according to their respective discrimination power.

Both unimodals as well as the bimodal one are trained and tested using the part of the DaFEx database which contains objects that are speaking and displaying facial expression simultaneously. Considering the performances of the unimodal systems, the results indicated that the one based on facial expression appears to be more successful compared to the one based on speech, while the bimodal one outperforms both. This notation is supported by a large body of evidence provided by theoretical studies on perception emotions by human encoders.

Further work needs to concentrate on several issues regarding both the facial-expression-based system and the bimodal one. Building an AAMs for new unseen persons demands proper annotation of some images of the considered subject. Up to now to the best of our knowledge all systems that utilize AAMs as feature extractor rely on the somewhat tedious manual annotation. The first simple solution to this problem might be a bootstrapping method that makes use of a combination of tracking of some fiducial facial points (e.g., irises and mouth center), knowledge-based methods, and the reconstruction error fed back from an already existing AAM.

Experiments carried out to evaluate the performance of the bimodal system indicate that firstly the bimodal system outperforms both unimodals; in other words, the cues of both modalities should be considered when aiming at an emotion analysis system that performs well in natural and social human-robot interaction. Furthermore, the results verified the suitability of our fusion scheme, in which the decision of each modality is fed after be weighted according to its discrimination power. Putting this fusion forward is proved by discussing the performance of each modality on each emotion.

An open issue regarding the used probabilistic fusion method is that the extension of the Bayesian network so that it includes a further variable that indicates if the interactant is speaking or not, i.e., indicates if the user is only displaying facial expression or she/he is speaking and displaying facial expression at the same time. This variable can be directly extracted from the speech-based analysis system. Regarding to this variable, new weights can be recomputed.

Both unimodal systems as well as the bimodal one performed well on the DaFEx database, while they delivered obviously lower recognition rates when they are employed

in real-life condition, event hough the number of emotion classes was bigger in the former case than the latter. It might be beneficial to have a more comprehensive real-life data captured by the robot directly for it to evaluate at will. Nevertheless, labeling such data presents another open challenge.

The robot's adaption to the needs of its interaction partner is not just simple mimicry. A future work towards complete human-human-like human-robot interaction is to build emotional profiles for both the robot and the interactant. An emotional profile of an interaction partner does not mean just recognizing the current emotional state of the interactant, but rather it could include his/her emotional behavior across multiple contexts, during several time periods, as well as her/his mood. Such a profile could provide a good basis for the robot to react the best according to these variables. The robot's emotional profile, in turn, could comprise an emotional-cognitive model, which would allow the robot to produce appraisals according to internal and external contexts. The latter can be presented as the information provided from the interactant's emotional profile, while the former could contain a combination of current stimuli with some existing schemas in its memory, representing a variety of information, past experiences, current goals and needs, and knowledge of the surrounding.

As a whole, the results suggest that facial expression and speech prosody provide information about the affective state of the interactant for the robot, and they are the most important input signals for natural human robot interaction. In addition, the results suggest that the recognition of each individual emotion is highly dependent on the used modality. Hence, both modalities should be considered in a joint manner, when a reliable emotion analysis system for natural, unconstrained, and real-life human-robot interaction is intended.

8 Appendix

8.1 Evaluation of Visual-Based System Using NN

A previous study is carried out by utilizing the facial-expression-based system proposed in Sec. 4.6. Instead of using SVM a rather simple nearest neighbor classifier is used. To classify a new face represented by the parameter vector c_i, assuming that the expression classes have a common covariance matrix, we measure the squared Mahalanobis distance $d_{\mathbf{M}}$ from c_i to each of the j estimated mean vector $\bar{c}_j$, $j \in$ (anger, disgust, fear, happiness, neutral, sadness, and surprise). the vector c_i is then assigned to the class of the nearest mean. mathematically we have computed $d_{mathbfM}(c_i, \bar{c}_j) = (c_i - \bar{c}_j)^t \mathbf{S}^{-1}(c_i - \bar{c}_j)$ for each new c_i and assigned it according to the class to which the vector has the lowest distance $d_{\mathbf{M}}$, where $\mathbf{S}$ is the Covariance matrix computed from some labeled data.

	Ang	Dis	Fea	Hap	Neu	Sad	Sur	Total
Ct	20.93	00.00	36.14	04.90	26.62	19.92	30.34	**19.84**
BB	54.24	52.52	55.73	59.95	88.49	63.27	55.89	**61.44**
LT	67.80	62.10	63.38	65.96	95.25	73.51	59.00	**69.57**
LW	83.04	71.34	63.77	65.68	91.38	75.93	54.76	**72.27**
GW	82.36	68.01	64.14	69.23	94.21	77.80	53.89	**72.80**

Table 8.1: Recognizing rates obtained from the facial-expression-based system exploiting person-dependent AAM. The system is evaluated on DaFEx database with a nearest neighbor classifier. Emotions are; Ang:Angry, Dis:Disgust, Fea, Fear, Hap:Happiness, Neu:Neutral, Sad:Sadness, and Sur:Suprise. Initialization methods; Ct:Centering, BB:Bounding Box, LT:Linear Transformation, LW:Linear Warping, and GW:Gaussian Warping.

As expected, Table. 8.1 shows that using person-dependent AMMs revealed better recognition rates as using person-independent AMMs, the reasons behind that are discussed in Sec. 6.3. However, comparing the performance of the system based on nearest neighbor classifier and that based on SVM indicates the advantage of the latter on the former. That is because the sensitivity of nearest neighbor classifier to the local structure of the data. This yields the classifier not to be able to form reliable neighborhoods and in consequence causes the classifier to fail on datasets with high level of sparsity, which is the case of our data.

	Ang	Dis	Fea	Hap	Neu	Sad	Sur	Total
Ct	31.09	9.76	12.85	56.22	00.00	10.03	14.89	**19.84**
BB	29.44	08.94	04.89	34.35	63.43	00.90	03.09	**20.72**
LT	33.08	12.20	01.15	39.95	53.98	03.60	10.83	**22.11**
LW	30.13	14.48	01.85	40.60	59.86	00.00	08.57	**22.21**
GW	30.16	12.60	00.62	41.16	56.86	00.00	10.83	**21.76**

Table 8.2: Recognizing rates obtained from the vision-based systems utilizing person-dependent Active Appearance model. The system is evaluated on DaFEx database with a nearest neighbor classifier. Emotions are; Ang:Angry, Dis:Disgust, Fea, Fear, Hap:Happiness, Neu:Neutral, Sad:Sadness, and Sur:Suprise. Initialization methods; Ct:Centering, BB:Bounding Box, LT:Linear Transformation, LW:Linear Warping, and GW:Gaussian Warping.

8.2 Home-Tour Scenario

In this scenario the robot is introduced to its new environment; *"a flat"* by a person without any knowledge of robotics. The interactant shows and names locations and objects which she believes are necessary for the robot to remember. In order for the robot to be able to allocate itself correctly the next time, it should have a robust mapping and localization methods in addition to reliable human-robot interaction abilities, to which belongs the understanding of interactant's emotions.

8.3 Object-Teaching Scenario

Like home-tour scenario, object-teaching scenario is also a learning task of BIRON, which is based on the perception abilities of BIRON and a reliable user-robot interaction. In this scenario an interactant shows a set of household subjects, such as bottle, cup, book, etc..., to BIRON and decides in a validation stage if BIRON has correctly learned the objects showed previously or not.

8.4 Notations

HCI: Human Computer Interaction

HRI: Human Robot Interaction

HHMs: Hidden Markov Models

NN: Nearest Neighbor Classifier

NNs: Neural Networks

SVMs: Support Vector Models

AAMs: Active Appearance Models

PCI: Principal Component Analysis

ICA: Independent Component Analysis

LDA: Linear Discriminant Analysis

EEG: Electroencephalograph

EMG: Electromyography

CPDs: Conditional Probability Distributions

CPTs: Conditional Probability Tables

BFFs: Basic Facial Features

KNN: K-Nearest Neighbor Classifier

Bibliography

[1] K. Anderson and P. McOwan. A real-time automated system for recognition of human facial expressions. *IEEE Transaction on Systems, Man, and Cybernetics*, 36:96–105, 2006.

[2] M. Arnold. *Emotion and Personality*. Columbia University Press, 1960.

[3] A. Asthana, J. Saragih, M. Wagner, and R. Goecke. Evaluating aam fitting methods for facial expression recognition. In *Int. Conf. on Affective Computing and Intelligent Interaction*. 2009.

[4] R. Banse and K. Scherer. Acoustic profiles in vocal emotion expression. *Personality and Social Psychology*, 70:614–636, 1996.

[5] M. Bartlett, B. Braathen, G. Littlewort, J. Hershey, I. Fasel, T. Marks, E. Smith, T. Sejnowski, and J. Movellan. Automatic analysis of spontaneous facial behavior. Technical report, Machine Perception Lab, Institute for Neural Computation, University of California, San Diego, 2001.

[6] M. Bartlett, M. Lades, and T. Sejnowski. Independent component representation for face recognition. In *Syposium on Electronic Imaging*. 1998.

[7] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Workshop on Computer Vision and Pattern Recognition*. 2003.

[8] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. 2005.

[9] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*. 2006.

[10] J. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Personality and Social Psychology*, 37:2049–2059, 1979.

[11] A. Battocchi, F. pianesi, and D. Goren-Bar. Dafex, a database of kinetic facial expression. In *ICMI05 Doctoral Spotlight and Demo Proceedings, 2005*. 2005.

[12] A. Battocchi, F. Pianesi, and D. Goren-Bar. A first evaluation study of a database of kinetic facial expressions (dafex). In *Proc. Int. Conf. Multimodal Interfaces*, pages 214–221. ACM Press, 2005.

[13] M. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25:23 – 48, 1997.

[14] M. Bradley, B. Cuthbert, and P. Lang. Picture media and emotion: Effects of sustained affective context. *Psychophysiology*, 33:662–670, 1996.

[15] M. Bradley and P. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 1:49–59, 1994.

[16] M. Bradley and P. Lang. Affective reactions to acoustic stimuli. *Psychophysiology*, 37:204–215, 2000.

[17] S. Brave and C. Clifford Nass. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, chapter Emotion in human-computer interaction. L. Erlbaum Associates Inc. USA, 2003.

[18] C. Breazeal. Emotion and social humanoid robots. *Human-Comuter Studies*, 59:119–155, 2003.

[19] Buehler. *Die Darstellungfunktion der Sprache*. Ungek. Nachdr. d. Ausg. Jena, Fischer, 1934.

[20] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. Int. Conf. Multimodal Interfaces*. 2004.

[21] G. Castellano, L. Kessous, and G. Caridakis. *Emotion recognition through multiple modalities: face, body gesture, speech*, chapter Affect and emotion in human-computer interaction, pages 92–103. Springer, New York, 2008.

[22] 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.

[23] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[24] Z.-J. Chuang and C.-H. Wu. Multi-modal emotion recognition from speech and text. *Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 2:45–62, 2004.

[25] J. Coan, J. Allen, and E. Harmon-Jones. Voluntary facial expression and hemispheric asymmetry over the frontal cortex. *Psychophysiology*, 38 cortex:912–925, 2001.

[26] M. Codispoti, M. Bradley, and P. Lang. Affective reactions to briefly presented pictures. *Int Journal of the Society for Psychophysiological Research*, 38:474–478, 2001.

[27] I. Cohen, N. Sebe, A. Garg, and L. S. Chen. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Undersatnding*, 91:160–187, 2003.

[28] J. Cohn and K. Schmidt. The timing of facial motion in posed and spontaneous smiles. *Jabor Wavelets, Multi-resolution and Information Processing*, 2:121–132, 2004.

[29] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Fifth European Conf. Computer Vision*. 1998.

[30] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *Fourth Int. Conf. on Automatic Face and Gesture Recognition*. 2000.

[31] L. Cosmides. Invariances in the acoustic expression of emotion during speech. *Journal of Experemental Psychology: Human Perception and Performance*, 9:864–881, 1983.

[32] R. Cowie and R. Cornelius. Describing the emotional states that are expressed in speech. *speech communication*, 40:5–32, 2003.

[33] R. Cowie, E. Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *ISCA Workshop Speech and Emotion*. 2000.

[34] A. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace, New York, 1999.

[35] C. Darwin. *The expression of emotions in man and animals*. University of Chicago Press, Chicago, 1965.

[36] B. de Gelder and J. Vroomen. Bimodal emotion perception: integration across separate modalities, cross-modal perceptual grouping or perception of multimodal events. *Cognition and Emotion*, 14:321–324, 2000.

[37] J. Donath. Mediated faces. In *4th Int. Conf. on Cognitive Technology*. 2001.

[38] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Tranaction on Pattern Analyis and Machine Intellegence*, 21:974–989, 1999.

[39] P. Ekman. *Human ethology: claims and limits of a new discipline*, chapter About brows: emotional and conversational signals., pages 169 – 222. Cambridge University Press, New York, 1979.

[40] P. Ekman. *Methods for measuring facial action*, chapter Handbook of methods in nonverbal behavior research, pages 45–90. Cambridge University Press, 1982.

[41] P. Ekman and W. Friesen. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*. Prentice Hall, 1975.

[42] P. Ekman, W. friesen, and P. Ellsworth. *Emotion in the Human Face*, chapter What Emotion Categories or Dimensions can observes Judge from Facial Behaviour, pages 39–55. Cambridge University Press, 1982.

[43] P. Ekman, W. Friesen, and S. Tomkins. Facial affect scoring technique, fast a first validity study. *Semiotica*, 3:37–38, 1971.

[44] P. Ekman and H. Oster. Facial expression of emotion. *Annual Reviews Psychology*, 30:527–554, 1979.

[45] I. A. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:757–763, 1997.

[46] S. Fagel. Emotional mcgurk effect. In *Proc. Int. Conf. on Speech Prosody*. Dresden, Germany, 2006.

[47] B. Fasel and J. Lüttin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36:259–275, 2003.

[48] J. Fernandez-Dols and M. Ruiz-Belda. Are smiles a sign of happiness gold medal winners at the olympic games. *Journal-of-Personalityand-Social-Psychology*, 69:1113–1119, 1995.

[49] N. Fragopanagos and J. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18:389–405, 2005.

[50] A. Fridlund. *The psychology of facial expression*, chapter The new ethology of human facial expressions, pages 103–129. Cambridge University Press, 1997.

[51] N. Frijda. *The Emotions: Studies in Emotion and Social Interaction*. Cambridge University Press, 1986.

[52] N. Frijda. *Emotions are functional, most of the time*, chapter The Nature of Emotion, page 112âĂŞ122. Oxford University Press, New York, 1994.

[53] J. Fritsch, M. Kleinehagenbrock, A. Haasch, S. Wrede, and G. Sagerer. A flexible infrastructure for the development of a robot companion with extensible hri-capabilities. In *Proc. ICRA*, pages 3419–3425. Barcelona, Spain, April 2005.

[54] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23:643–660, 2001.

[55] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. 23:1080–1093, 2005.

[56] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. 2006.

[57] S. Gunn and M. Nixon. Snake head boundary extraction using global and local energy minimisation. In *13th Int. Conf. on Pattern Recognition*. 1996.

[58] M. Hanheide, S. Wrede, C. Lang, and G. Sagerer. Who am i talking with? a face memory for social robots. In *IEEE Int. Conf. on Robotics and Automation*. 2008.

[59] J. Healey. *Wearable and Automotive Systems for Affect Recognition from Physiology*. Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institut of Technology, 2000.

[60] F. Hegel, T. Spexard, T. Vogt, G. Horstmann, and B. Wrede. Playing a different imitation game: Interaction with an empathic android robot. In *Proc. Int. Conf. Humanoid Robots*, pages 56–61. 2006.

[61] U. Hess. The communication of emotion. *Emotions, Qualia, and Consciousness*, pages 397–409, 2001.

[62] J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke. Distinguishing deceptive from non-deceptive speech. In *9th European Interspeech*. 2005.

[63] C. Huang, H. Ai, , Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *IEEE Transaction on Pattern Analysis and Macine Intellegence*, 29:671–686, 2007.

[64] R. Huang and C. Ma. Toward speaker independent real time affect detection system. In *International Conference on Pattern Recognition*. 2006.

[65] X. Huang, S. Z. Li, and Y. Wang. Statistical learning of evaluation function for asm/aam image alignment. In *ECCV Workshop BioAW*. 2004.

[66] E. Hudlicka. To feel or not to feel: The role of affect in human-computer interaction. *Human-Computer Studies*, 59:1–32, 2003.

[67] C. Izard. *Human Emotions*. Plenum Press, 1977.

[68] C. Izard and L. Dougherty. *Measuring Emotions in Infants and Children*, chapter Two Complementary Systems for Measuring Facial Expressions in infants and Children. Cambridge University Press, 1982.

[69] Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and raealtime monitoring human fatigue. *IEEE Transaction on Systems, Man and Cybernetics*, 36:862–875, 2005.

[70] R. E. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Int. Conf. Computer Vision and Pattern Recognition*. 2004.

[71] R. E. Kaliouby and P. Robinson. *Real-Time Vision for Human-Computer Interation*, chapter Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures, pages 180–200. Springer, USA, 2005.

[72] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *4th IEEE Int Conf on Automatic Face and Gesture Recognition*. 2000.

[73] A. Kapoor, W. Burleson, and R. Picard. Automatic prediction of frustration. *Int. Journal. Human-Computer Studies*, 65:724–736, 2007.

[74] K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis, and S. Kollias. *Facial Expression and Gesture Analysis for Emotionally-Rich Man-Machine Interaction*, chapter 3D modeling and animation: synthesis and analysis techniques for the human body. Idea Group, 2004.

[75] J. Kim. *Robust Speech Recognition and Understanding*, chapter Bimodal Emotion Recognition Using Speech and Physiological Changes, pages 265–280. I-Tech Education and Publishing, 2007.

[76] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.

[77] P. Kleinginna and A. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, Volume 5,4:345–379, 1981.

[78] C. Kotropoulos and I. Pitas. Rule-based face detection in frontal views. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. 1997.

[79] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transaction on Image Processing*, 16:172–187, 2007.

[80] P. Lang, M. Greenwald, M. Bradley, and A. Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30:261–273, 1993.

[81] J.-J. Lien. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*. Ph.D. thesis, The Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1998.

[82] C. Lisetti and C. LeRouge. Affective computing in tele-home health. In *Proceedings of the 37th Hawaii International Conference on System Sciences*. 2004.

[83] G. Littlewort, M. Bartlett, and K. Lee. Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Ninth ACM Int. Conf. on Multimodal Interfaces*. 2007.

[84] C. Liua, K. Conna, N. Sarkarb, and W. Stonec. Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder. *Human-Computer Studies*, 66:662–677, 2008.

[85] S. Lucey, A. B. Ashraf, and J. Cohn. *Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face*, chapter Face Recognition, pages 275–286. I-Tech Education and Publishing, 2007.

[86] I. Lütkebohle, F. Hegel, S. Schulz, M. Hackel, B. Wrede, S. Wachsmuth, and G. Sagerer. The bielefeld anthropomorphic robot head, flobi. In *ICRA*. 2010, in press.

[87] M. Lyons and S. Akamatsu. Coding facial expressions with gaborwavelets. In *3rd IEEE International Conference on Automatic Face and Gesture Recognition*. 1998.

[88] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transaction on Pattern Analysis and Machine Intellegence*, 12:1357–1362, 1999.

[89] R. Mandryk and S. Atkins. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *Int. J. Human-Computer Studies*, 65:329–347, 2007.

[90] M. Mansoorizadeh and N. M. Charkari. Bimodal person-dependent emotion recognition: comparison of feature level and decision level information fusion. In *HCI-HRI workshop*. 2008.

[91] D. Massaro and P. Egan. Perceiving affect from the voice and the face. *Psychonomoic Bulletin and Review*, 3:215–221, 1996.

[92] Y. Matsusaka, T. Tojo, and T. Kobayashi. Conversation robot participating in group conversation. *IEICE Transactions on Information and Systems*, E86-D:26–36, 2003.

[93] R. McFarland. Relationship of skin temperature changes to the emotions accompanying music. *Biofeedback and Self-Regulation*, Vol. 10:255–267, 1985.

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

[95] A. Mehrabian and J. Russell. *An Approach to Environmental Psychology*. MIT Press, Cambridge U, 1974.

[96] P. Michel and R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. 2003.

[97] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. *Automatic Systems for the Identification and Inspection of Humans, SPIE*, 1994.

[98] O. Mowrer. *Learning theory and behavior*. Wiley, New York, 1960.

[99] C. Nass, J. Steuer, and E. Tauber. Computers are social actors. In *Proceeding of CHI Conference on Human factors in Computing systems*. 1994.

[100] K. Oatley. Towards a cognitive theory of emotion. *Cognitive and Emotion*, 1:29–50, 1987.

[101] K. Oatley and J. Jenkins. *Understanding emotions*. Wiley-Blackwell, 1996.

[102] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[103] A. Ortony and T. Turner. What's basic about basic emotions. *Psychological Review*, 97, 3:315–331, 1990.

[104] T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using hmm. In *Second Int. Conf. on Automatic Face and Gesture Recognition*. 1998.

[105] C. Padgett and G. Cottrell. Representing face images for emotion classification. *Advances in Neural Information Processing System*, 9:894–900, 1997.

[106] M. Paleari and C. Lisetti. Toward multimodal fusion of affective cues. In *Proc. ACM int. workshop on Human-centered multimedia*, pages 99–108. ACM, New York, NY, USA, 2006.

[107] M. Pantic and M. Bertlett. *Machine Analyis of Facial Expressions*, chapter Face Recognition, pages 377–416. I-Tech Education and Publishing, 2007.

[108] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments form face profile image sequences. *IEEE Transaction on Systems, Man, and Cybernetics*, 36:433–449, 2006.

[109] M. Pantic and J. Rothkrantz. Expert system for automatic analysis of facial expressions. *Inmage and Vision Computing*, 11:881–905, 2000.

[110] M. Pantic and J. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transaction on Systems, Man, and Cybernetics*, 34:1449–1461, 2004.

[111] M. Pantic, M. Valstar, R. rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE*. 2005.

[112] T. Partala. *Affective Information in Human-Comuter Interaction*. Ph.D. thesis, Department of Computer Sciences University of Tampara, 2005.

[113] paul Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Palo Alto : Consulting Psychologists Press*, 1978.

[114] R. Picard. *Affective Computing*. MIT Press, 1997.

[115] R. Plutchik. *Emotion: Theory, research, and experience*, chapter A general psychoevolutionary theory of emotion, pages 3–33. Random House, New York., 1980.

[116] Y. Qi, C. Reynolds, and R. Picard. The bayes point machine for computer-user frustration detection via pressuremouse. In *Proceeding of workshop on Perceptive User Inteface*. 2001.

[117] A. Rabie, C. Lang, M. Hanheide, M. Castrillón-Santana, and G. Sagerer. Automatic initialization for facial analysis in interactive robotics. In *Proc. Int. Conf. Computer Vision Systems*. Santorini, Greece, May 2008 2008.

[118] A. Rabie, T. Vogt, M. Hanheide, and B. Wrede. Evaluation and discussion of multimodal emotion recognition. In *ICCEE*. 2009.

[119] L. Rabiner. A tutorial on hidden markov models and selected applications in speech processing. *IEEE*, 77:257–285, 1989.

[120] C. Randolph. *The Science of Emotion*. Engliewood Cliffs, Prentice Hall, 1996.

[121] B. Reeves and C. Nass. *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press New York, 1996.

[122] J. Rong, Y.-P. P. Chen, M. Chowdhury, and G. Li. Acoustic features extraction for emotion recognition. In *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*. 2007.

[123] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20:23–38, 1998.

[124] J. Russel. A circumplex model of affect. *Personality and Social Psychology*, 39:1161–1178, 1980.

[125] J. Russel. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin [0033-2909] Russell yr:1994 vol:115 pg:102*, 115:102–141, 1994.

[126] J. Russel, J.-A. Bachorowski, and J. Fernandez-Dols. Facial and vocal expression of emotion. *Annu. Rev. Psychol*, 54:329–349, 2003.

[127] J. Russell and J. Fernandez-Dols. *The Psychology of Facial Expression*, chapter What does a facial expression mean, pages 3–30. University of Cambridge Press, Cambridge, UK, 1997.

[128] S. Russell and P. Norvig. *Künstliche Intelligenz*. Pearson Education, Germany, 2003.

[129] A. Samal and P. Iyengar. Automatic recognition and analysis of human face and facial expression, a survey. *Pattern Recognition*, 25:65–77, 1992.

[130] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *2nd IEEE Workshop on Applications of Computer Vision*. 1994.

[131] A. Savran, N. Alyüz, H. Dibeklioglu, O. Celiktutan, B. Gökberk, B. Sankurand, and L. Akarun. Bosphorus database for 3d face analysis. In *The First COST 2101 Workshop on Biometrics and Identity Management (BIOID)*. 2008.

[132] K. Scherer. Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communication*, 1:1–98, 1987.

[133] K. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227 – 256, 2003.

[134] K. Scherer, R. Banse, and H. Wallbott. Emotion inference from vocal expression correlate across languages and cultures. *Cross-Cultural Psychology*, 32:76–92, 2001.

[135] K. Scherer, R. Banse, H. Wallbott, and T. Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15, 2, 1991.

[136] J. Schwartz. Why the FLMP should not be applied to McGurk data .. or how to better compare models in the bazesian framework. In *Proc. Int. Conf. Audio-Visual Speech Processing*, pages 77–82. 2003.

[137] N. Sebe, M. Lew, I. Cohen, A. Garg, and T. Huang. Emotion recognition using a cauchy naive bayes classifier. 2002.

[138] N. Sebe, M. Lew, I. C. nad Yafei Sun, T. Gevers, and T. Huang. Authentic facial expression analysis. In *Int. Conf. Automatic Face and Gesture Recognition*. 2004.

[139] H. Siegwart and K. Scherer. Acoustic concomitants of emotional expression in operatic singing: The case of lucia in ardi gli incensi. *Journal of Voice*, 9:249–260, 1995.

[140] L. D. Silva and P. Chi. Bimodal emotion recognition. In *fourth IEEE int. conf. on automatic face and gesture recognition*. 2000.

[141] L. D. Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multi-modal information. In *IEEE Int. Conf. on Information, Communications and Signal Processing*. 1997.

[142] S. Sirohey. Human face segmentation and identification. Technical report, Maryland University, 1993.

[143] M. Song, M. You, N. Li, and C. Chen. A robust multimodal approach for emotion recognition. *Neurocomputing*, 71:1913–1920, 2008.

[144] T. Spexard, M. Hanheide, and G. Sagerer. Human oriented interaction with an anthropomorphic robot. *IEEE Transactionson Robotics Special Issue on Human Robot Interaction*, 23:852–862, 2007.

[145] Q. Summerfield. Use of visual information in phonetic perception. *Phonetica*, 36.:314–331, 1979.

[146] K. Takahashi. Remarks on emotion recognition from bio-potential signals kazuhiko takahashi. In *2nd International Conference on Autonomous Robots and Agents*. 2004.

[147] A. L. Thomaz, M. Berlin, and C. Breazeal. An embodied computational model of social referencing.

[148] Y. Tian, T. Kanade, and J. F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. 2002.

[149] Y.-L. Tian, T. Kanade, and J. F. Cohn. *Facial Expression Analysis*, chapter Handbook of Face Recognition, pages 247–275. Springer, 2005.

[150] S. Tomkins. *Approaches to Emotions*, chapter Affect Theory, pages 163–195. Hillsdale, Erlbaum, 1984.

[151] M. Tscherepanow, M. Hillebrand, F. Hegel, B. Wrede, and F. Kummert. Direct imitation of human facial expressions by a user-interface robot. In *9th IEEE Int. Conf on Humanoid Robots*. 2009.

[152] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Computer Vision and Pattern Recognition Workshop*. 2006.

[153] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection from face video. In *IEEE Int. Conf. on Systems, Man, and Cybernetics*. 2004.

[154] V. Vapnik. An overview of statistical learning theory. *IEEE Transaction on Neural Networks*, 19:988–999, 1999.

[155] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[156] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proc. of IEEE Int. Conf. on Multimedia & Expo*. Amsterdam, The Netherlands, July 2005.

[157] T. Vogt, E. André, and N. Bee. Emovoice — A framework for online recognition of emotions from voice. In *Proc. Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Irsee, Germany, June 2008.

[158] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*. 2005.

[159] H. Wang, H. Prendinger, and T. Igarashi. Communicating emotions in online chat using physiological sensors and animated text. In *Conference on Human Factors in Computing Systems*. 2004.

[160] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. 2006.

[161] M. Wang, Y. Iwai, and M. Yachida. Expression recognition from time-sequential facial images by use of expression change model. In *Second Int. Conf. on Automatic Face and Gesture Recognition*. 1998.

[162] T. Wehrle and S. Kaiser. *Emotion and Facial Expression*, chapter Affective Interaction, pages 49–63. Springer Berlin, 2006.

[163] B. Weiner and S. Graham. *Emotions, cognition, and behavior*, chapter An attributional approach to emotional development, pages 167–191. Cambridge University Press, New York, 1984.

[164] Z. Wen and T. Huang. Capturing subtle facial motions in 3d face tracking. In *Ninth IEEE International Conference on Computer Vision*. 2003.

[165] G. Yang and T. Huang. Human face detection in complex background. *Pattern Recognition*, 27:53–63, 1994.

[166] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transaction on Pattern Analysis and Macine Intellegence*, 24:34–58, 2002.

[167] K. Yow and R. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15:713–735, 1997.

[168] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. Huang. One-class classification for spontaneous facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*. 2006.

[169] Z. Zeng, Y. Hu, Y. Fu, T. S. Huang, G. I. Roisman, and Z. Wen. Audio-visual emotion recognition in adult attachment interview. In *Proc. Int. Conf. on Multimodal Interfaces*, pages 139–145. ACM, New York, NY, USA, 2006.

[170] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. In *ICMI*. 2007.

[171] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transaction on Pattern Analysis and Macine Intellegence*, 31:39–58, 2009.

[172] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. S. Huang, D. Roth, and S. Levinson. Bimodal hci-related affect recognition. In *ICMI*. 2004.

[173] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transaction on Pattern Analysis and Machine Intellegence*, 27:699–714, 2005.

[174] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Second International Conference on Automatic Face and Gesture Recognition*. 1998.

[175] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Systems CSUR*, 35:399–458, 2003.