

Imperial College of Science, Technology and Medicine
Department of Computing

Timing is everything

**A spatio-temporal approach to the analysis of facial
actions**

Michel François Valstar

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College, February 2008

Abstract

This thesis presents a fully automatic facial expression analysis system based on the Facial Action Coding System (FACS). FACS is the best known and the most commonly used system to describe facial activity in terms of facial muscle actions (i.e., action units, AUs). We will present our research on the analysis of the morphological, spatio-temporal and behavioural aspects of facial expressions. In contrast with most other researchers in the field who use appearance based techniques, we use a geometric feature based approach. We will argue that that approach is more suitable for analysing facial expression temporal dynamics. Our system is capable of explicitly exploring the temporal aspects of facial expressions from an input colour video in terms of their onset (start), apex (peak) and offset (end).

The fully automatic system presented here detects 20 facial points in the first frame and tracks them throughout the video. From the tracked points we compute geometry-based features which serve as the input to the remainder of our systems. The AU activation detection system uses GentleBoost feature selection and a Support Vector Machine (SVM) classifier to find which AUs were present in an expression. Temporal dynamics of active AUs are recognised by a hybrid GentleBoost-SVM-Hidden Markov model classifier. The system is capable of analysing 23 out of 27 existing AUs with high accuracy.

The main contributions of the work presented in this thesis are the following: we have created a method for fully automatic AU analysis with state-of-the-art recognition results. We have proposed for the first time a method for recognition of the four temporal phases of an AU. We have build the largest comprehensive database of facial expressions to date. We also present for the first time in the literature two studies for automatic distinction between posed and spontaneous expressions.

Acknowledgements

This thesis would not have been possible without my supervisor, Dr Maja Pantic. I would like to thank her for her tireless enthusiasm for my research and the many hours that we spent discussing my work and brainstorming for new ideas. From the very beginning she took me serious as a fellow scientist. She always allowed me to use her encyclopaedic knowledge of the facial expression recognition field, which was invaluable to me. Over the years she has become to me and the group more than just a supervisor. Also many thanks to Prof. Daniel Rueckert, my second supervisor, for always taking the time to answer whatever questions I had.

Dr Yiannis Patras was my second supervisor during my Masters project at Delft University of Technology. Ever since, we've kept in touch and he's been an incredibly valuable source of information. Especially regarding the problem of computer vision, his knowledge has helped me many times.

Prof. Jeffrey Cohn was in my Masters of Science committee. During the summer of 2006 we were his guest at Carnegie Mellon University/University of Pittsburgh. He has been a great host and we have had a fruitful research cooperation ever since. I am very grateful for the opportunities that he has given me.

This has been a very heavy year for me and my family. But we stucked together and pulled it through. Mamma, David and Judith, thank you so much for keeping up with my terrible mood swings this year and for giving me the energy and motivation to see it through. I wouldn't have reached this point without your support.

A special thanks should go to Sjoerd and Anna, who provided me with a port of call in London whenever I had troubles, be it with my research, being away from my home country or just personal issues. You were most patient with me and I love you for all that you've done for me.

As the Roman poet Juvenal said, there is a healthy mind in a healthy body. I probably wouldn't have stayed sane in the past year without Capoeira. I would like to thank my Mestre Ponciano, Contra-Mestre Casquinha and everybody from Cordao de Ouro London for giving me a new family in London, and Marquinhos Marmotta and everybody from Liberdade for still regarding me as a family, welcoming me with open arms every time I came to visit them in Delft.

And finally I would like to thank all the people in the college with whom we've always had our daily lunch rituals. I am sure we've got the best office in Imperial! Thank you Antonis, Stavros, Margarita, Emma, Uri, Theo, Kanwal, Sander, Edwin, Daniel, Gerardo, Brais and Monica.

Dedication

To my father, de zeerweledelgeleerde heer Dr. M.C. Valstar (February 6 1937 - August 13 2007),
who always inspired me to learn and understand.

‘His face bespoke his soul‘

Voltaire

‘Timing is everything.‘

Tommy Shaw

Contents

Abstract	i
Acknowledgements	ii
1 Introduction	1
1.1 Contributions	5
1.2 Application areas	6
1.3 Publications	9
2 Facial Expression Analysis	11
2.1 Facial affect theory and facial action coding	13
2.1.1 Ekman’s Six Basic Emotions	13
2.1.2 Plutchik’s Emotion Wheel	15
2.1.3 Russell’s Circumplex model	17
2.1.4 Facial Action Coding System (FACS)	18
2.1.5 MPEG-4	25
2.2 Automatic Facial Expression Analysis	26
2.2.1 Geometric feature based FACS detectors	29
2.2.2 Temporal dynamics analysis	34

3	Facial expression databases	38
3.1	Facial expression database characteristics	39
3.1.1	Desirable data characteristics	39
3.1.2	Data collection issues	43
3.1.3	Database accessibility and usability	43
3.2	Overview of existing databases	44
3.3	Cohn-Kanade database	45
3.4	MMI Facial expression database	47
4	Facial feature point detection	51
4.1	Face detection	54
4.2	ROI detection	56
4.3	Gabor Feature Extraction	58
4.4	Facial Point Detection using GentleBoost	61
4.5	Evaluation	64
5	Feature extraction	66
5.1	Facial feature point tracking	66
5.1.1	Condensation algorithm	68
5.1.2	Particle filtering with Factorised likelihoods	70
5.1.3	Rigid and morphological observation models	71
5.1.4	Transition model and priors	73
5.2	Registration of tracking data	75
5.2.1	Intra-sequence registration	76
5.2.2	Inter-sequence registration	76

5.2.3	Smoothing the tracking data	77
5.3	Mid-level parameter extraction	78
5.3.1	Single frame based features	78
5.3.2	Inter frame based features	79
5.3.3	Local time parameterised features	80
6	Action Unit activation detection	82
6.1	Performance measures	83
6.2	Evaluation procedures	85
6.3	Action Unit activation detection	87
6.3.1	Feature Selection	87
6.3.2	Support Vector Machine Classification	92
6.3.3	Cascade Support Vector Machines	93
6.4	Evaluation of AU activation detection	96
6.4.1	Frame-based AU detection results	96
6.4.2	Event coding using AU frame activation information	107
7	Action Unit temporal analysis	110
7.1	Temporal analysis per frame	111
7.2	Temporal model using temporal dynamics grammatical rules	113
7.3	Temporal model using Hidden Markov Models	114
7.4	Feature set evaluation	122
7.5	Event detection using AU temporal analysis	124

8	Emotion recognition	128
8.1	Emotion recognition: conscious vs. unconscious reasoning approaches	128
8.2	One-step emotion recognition	130
8.3	Two-step emotion recognition	131
8.4	Experiments	131
8.5	Application: Emotionally Aware Painting Fool	134
9	Applications to human facial behaviour understanding	137
9.1	Related work in automatic spontaneous facial expression analysis	138
9.2	Posed vs. spontaneous brow actions: automatic detection	140
9.2.1	Mid-level feature parameters	141
9.2.2	Classification strategy	145
9.2.3	Evaluation	147
9.3	Posed vs. spontaneous smiles: multi-cue automatic detection	150
9.3.1	Tracking	153
9.3.2	Temporal segmentation	155
9.3.3	Fusion strategies	157
9.3.4	Evaluation	162
10	Conclusion	169
10.1	Summary of Thesis Achievements	170
10.1.1	Feature definitions	170
10.1.2	Fully automatic AU analysis	171
10.1.3	Temporal dynamics analysis	171
10.1.4	Emotive expression recognition	172

10.1.5	MMI-Facial Expression Database	173
10.1.6	Distinguishing posed from spontaneous facial actions	173
10.2	Future Work	174

Bibliography		175
---------------------	--	------------

List of Tables

2.1	Facial muscles used in Action Units.	23
2.2	Upper facial Action Units.	24
2.3	Lower facial Action Units.	24
2.4	Action Units belonging to neither the upper nor the lower facial area.	24
3.1	Existing facial expression databases	44
3.2	Characteristics of facial expression databases	45
4.1	Definition of the 20 facial points used in our system.	53
4.2	Characteristic Facial Point detection results for 300 samples from the Cohn-Kanade database	65
4.3	Characteristic Facial Point detection results for 244 samples from the MMI-Facial Expression Database. e_p is the average relative error, measured in units of Interocular Distance D_I	65
6.1	Performance of frame-based AU activation detection for various feature sets on MMI database.	97
6.2	First four features selected by GentleBoost for the activation detection of AUs in the MMI database.	99
6.3	Subject independent cross validation results for AU activation detection per frame on 244 examples from the MMI-Facial Expression Database	101

6.4	Subject independent cross validation results for AU activation detection per frame on 153 examples from the Cohn-Kanade Database	103
6.5	F1-measure for cross-database AU detection per frame (second and third column) and per video (last two columns). The system was either trained on 244 examples from the MMI-Facial Expression Database and tested on 153 examples from the Cohn-Kanade Database or trained on Cohn-Kanade and tested on MMI.	105
6.6	Subject independent cross validation results for AU event detection using adaptive thresholds on 244 examples from the MMI Facial Expression Database. The last column shows the results for fixing $\theta = 0$	109
6.7	Subject independent cross validation results for AU activation event detection using adaptive thresholds on 153 examples from the Cohn-Kanade Database. The last column shows the results for fixing $\theta = 0$	109
7.1	Classification accuracy of multiclass gentleSvm at distinguishing the four temporal phases neutral, onset, apex and offset, measured per frame in terms of F1-measure. . .	113
7.2	Classification accuracy of multiclass gentleSvm classifier followed by a filtering process using expression grammatical rules for distinguishing the four temporal phases neutral, onset, apex and offset, measured per frame, in terms of F1-measure.	115
7.3	Classification accuracy of hybrid SVM-HMM classifier at distinguishing the four temporal phases neutral, onset, apex and offset, measured per frame, in terms of F1-measure.	118
7.4	Performance of various feature set combinations for the problem of recognising the temporal phases of Action Units. The fourth to seventh column show the average values over all AUs, per temporal phase.seventh column show the average values over all AUs, per temporal phase. The second and third columns show the average values over all temporal phases and all AUs.	124
7.5	Most important features for distinguishing pair-wise between the four temporal segments neutral, onset apex and offset.	125
7.6	Comparison of the various event detection methods.	126
7.7	Subject independent cross validation results for AU activation event detection after identification of the temporal segments of AUs. Results are for 244 examples taken from the MMI Facial Expression Database	127

8.1	Rules for mapping AUs to emotions, according to the FACS investigators guide. A B means ‘either A or B’	129
8.2	Results of one-step emotion recognition, <i>clr</i> is classification rate, <i>rec</i> is recall and <i>pr</i> is precision: A) classification of features to emotions by a multi-class SVM, B) one-step emotion classification without feature selection.	132
8.3	Results of two-step emotion recognition, <i>clr</i> is classification rate, <i>rec</i> is recall and <i>pr</i> is precision: A) classification of manually labelled AUs to emotions by rules, B) automatically detected AUs classified to emotions by rules, C) Neural Networks classifying automatically detected AUs into emotions	132
9.1	Mid-level feature parameters that GentleBoost selected as the most informative for determining whether a detected temporal segment <i>d</i> of an activated AU has been displayed spontaneously or not. The 1st column lists the 9 relevant classes, the 2nd column lists the total number of mid-level parameters selected by GentleBoost for the relevant class, and the 3rd column lists the three most informative of the selected parameters defined by equations (9.1)-(9.11)	148
9.2	Correct classification rates attained by 9 RVMs using the mid-level feature parameters that the Gentle Boost selected as the most informative for the classification problem in hand, i.e., determining whether a detected temporal segment of an activated AU has been displayed spontaneously or not.	149
9.3	Final classification results achieved by the probabilistic decision function defined by equation (9.13) for the entire brow actions shown in an input face image sequence. For the purposes of computing the recall and precision, the spontaneous class was considered the target class.	149
9.4	Description of the three late fusion criteria used: sum, product and weight.	162
9.5	Selected low-abstraction features to distinguish posed from spontaneous smiles.	162
9.6	Classification, recall, precision rates and F1-measure for the different fusion strategies employed. Performance measures are computed per video.	164
9.7	Comparison of performance measures for the different visual cues separately and fused. Performance measures are computed per frame.	165

9.8	Matrix of statistical significant different classification rates. Roman indices relate to the visual cue combinations listed in table 9.7. A 1 indicates statistically significantly different results.	165
9.9	Classification rates for the specialised classifiers used in late fusion. There is no temporal phase associated with the concurrency classifier.	166
9.10	Selected high-abstraction features to distinguish posed from spontaneous smiles. . . .	167

List of Figures

2.1	Plutchik’s emotion wheel (left) and the accompanying emotion cone (right).	16
2.2	Russell’s circumplex model of emotions, showing the emotion space spanned by the bipolar continui <i>valence</i> and <i>arousal</i>	17
2.3	Some examples of Action Units. Action Units are atomic facial muscle actions described in FACS.	18
2.4	The facial muscles seen from the side (image taken from Gray’s Anatomy)	19
2.5	Upper face muscles used to produce facial expressions (image taken from Gray’s Anatomy).	20
2.6	Upper face muscles used to produce facial expressions (image taken from Gray’s Anatomy).	20
2.7	Illustration of the elevator palpebrae superior, used to raise the upper eyelids (AU5, image taken from Gray’s Anatomy).	21
3.1	Examples of static frontal-view images of facial expressions in the MMI Facial Expression Database	48
3.2	Examples of apex frames of dual-view image sequences in MMI Facial Expression Database	48
4.1	Fiducial facial point model (left) and fiducial facial points annotated on an image of a neutral face from the MMI Facial Expression Database	54
4.2	Outline of the fiducial facial point detection system.	55
4.3	Positive and negative examples for training the classifier for the right inner eye corner. The big white square represents the 9 positive examples. Eight negative examples have been randomly picked near the positive examples and another 8 are randomly chosen from the remainder of the region of interest.	62

4.4	Examples of accurately detected facial points.	64
5.1	Results from the fiducial facial point tracker, which uses Particle Filtering with Factorized Likelihoods. 20 fiducial facial points were tracked in 332 sequences of up to 200 frames taken from the MMI Database and the Cohn-Kanade database. The top three rows are from the MMI database and the bottom two rows from the Cohn-Kanade database. All rows start with a neutral expression. The	69
5.2	The assumed transition model. α_i and α_i^- are the current and the previous state of a facial point i and y_i and y_i^- are the current and the previous observations. Each facial point's state can be represented as a simple Markov chain in temporal domain whilst in each time instant different facial points may be related. The dashed lines represent those interdependencies.	74
5.3	Noise reduction by applying a temporal filter to two features relevant for detection of Action Unit 1 (inner brow raise). The x axis represents the x-position of point D (right inner brow) and the y-axis represents its y-position. Circles are examples from frames where the AU was not active, squares are from frames where the AU was at its peak. The left figure shows the unfiltered features, while the right figure clearly shows a reduction in noise and clearer spatio-temporal patterns.	77
6.1	Overview of the fully automatic AU analysis system.	83
6.2	Performance of GentleBoost with different initialisations of the weights, measured in classification rate and F1-measure.	90
6.3	The performance of AU-activation recognition for different fixed numbers of features selected by GentleBoost.	91
6.4	Schematical overview of the Cascade SVM	94
6.5	Effect of the number of cascade loops on the classification performance.	95
6.6	Effect of the number of feedback loops on the classification performance.	96
6.7	Evaluation of the effect of the number of positive examples in a dataset.	104

7.1	Scatter plots of pairs of the most informative mid-level feature parameters selected by GentleBoost for: (left) the onset temporal segments of AU1, (mid) the apex temporal segments of AU1 and (right) the offset temporal segments of AU1. Crosses denote spontaneous brow data while squares denote data of deliberately displayed brow actions.	112
7.2	An example of AU temporal phase recognition for AU25. The solid line shows the true values of the phase labelling per frame and the dotted line the prediction by the SVM-HMM. Horizontal lines depict either a neutral or an apex phase, an upward slope signifies an onset phase and a downward slope an offset phase.	117
7.3	Comparison of the classification results shown per temporal phase (onset, apex, offset and neutral). The results shown are the average over all 22 AUs.	119
7.4	Phase duration error of the detected temporal phases onset, apex and offset, and the entire facial action. Results are averaged per AU, and measured frames.	120
7.5	Relative phase duration of the detected temporal phases onset, apex and offset, and the entire facial action. Results are averaged per AU, and measured frames.	121
7.6	Average number of frames a phase starts late (positive values) or early (negative values), for the temporal phases onset, apex and offset, per AU.	122
7.7	Breakdown of how often a temporal phase starts early, late, or on time.	123
8.1	Overview of the automatic AU and emotion detection system: (a) input image sequence with tracked facial points, (b) tracking results, (c) the four most important features for recognition of AU2 shown over time, (d) GentleBoost is used to select the most important features, (e) which are subsequently fed to (mc-)SVMs, (f) for two step approach: emotion detection from AUs by Artificial Neural Networks or a rulebase. . .	132
8.2	Apex frame captured of the author showing his feelings during the Machine Intelligence Competition 2007.	135
8.3	Portrait of the author showing his feelings during the Machine Intelligence Competition 2007.	136
9.1	Illustration of the tracking procedure and the points used to obtain the tracking data: (a-c) for the head and shoulder modalities, and (d) for the face modality.	154
9.2	Tracked points $T_{f1} \dots T_{f12}$ of the face and tracked points $T_{s1} \dots T_{s5}$ of the shoulders. .	155

9.3 Mean velocity of the right mouth corner in x-direction during onset (x-axis) vs. mean velocity of the right mouth corner in x-direction during offset (y-axis). Crosses denote spontaneous smiles. 167

Chapter 1

Introduction

The computer is a tool that humankind has been using for only a short time, and the way we use it is constantly and rapidly changing. In the past, we used computers to speed up our work or, when they became small and cheap enough to enter our homes, to provide entertainment. Nowadays computers are no longer optional but essential to the functioning of our society. They control such important tasks as stocking our supermarkets, guiding trains, traffic lights and they monitor the quality of our water supply. The functioning of the computers performing these tasks is hidden to the general public. This is because we do not interact with these computers directly. Instead the computer is *indirectly aware* of our actions without us even noticing it (for instance, a system counts how many tubes of toothpaste are sold) and we only perceive the outcome of the computer's actions (there's always toothpaste in the store).

Just about every machine we own has some form of computational intelligence by the use of a micro-processor. Our mobile phones, cars, stereos, media players, thermostats and washing machines are all examples of this. But for all their powerful computing capabilities, we are painstakingly aware of our tedious interaction with them. Often we feel that those machines are actually quite stupid, and don't understand what we really want or need. Your 'intelligent' thermostat using temperature sensors fails to notice that you're actually feeling cold and your media player does not understand that you are not in the mood right now for that sad, slow music. What is missing at this moment is a breakthrough in the way machines sense human actions to understand their intentions. Only then can a natural, human-centred way of interaction with our computerised environment be created.

In the future, we will probably interact even more frequently with computers and other intelligent

machines, for an increasing number of everyday tasks. Futurists envision our environment equipped with a plethora of invisible sensors that will anticipate every need of a human being. Our homes, our means of transportation, the place where we work, will all intelligently interact with us. Social robots will guide us in museums and help in our households. This vision of the future is often referred to as *ubiquitous computing* [Weiser, 1991], *ambient intelligence* [Aarts, 2005], or *human computing* [Pantic et al., 2006].

Due to the nature of these smart environments and the need to interact with invisible, intelligent, embedded machines, an era of novel interfacing will arise. Machines will no longer have a well-defined physical presence and by their capabilities of interconnection will not be confined to one central location. This, in turn, means that a user will no longer be required to use one spatially fixed console. Instead, interaction with the machine(s) will preferably be through many interconnected, unobtrusive, remotely embedded sensors, shifting the user input mainly to remote sensing.

This remote sensing could free us from the primitive, rigid interaction with machines. Instead of using a keyboard and a mouse, a person could issue commands using his/her voice, facial expression and gestures. The machine will *read* human motion, identity, posture, gait, hand gestures, facial expression etc. This is called *human sensing* [Pantic et al., 2006]. The smart environment can then give feedback by executing the perceived command or, if it is uncertain about the command given, by asking for confirmation or extra information. Feedback could be provided through a synthetic voice, visual output on screens, or, in a more abstract fashion, by changing environmental variables such as lighting, temperature or volume of music. This will provide the natural way of interaction with machines that we aim for.

Besides reading instantaneous commands the smart environment could also record and analyse long-term behaviour of humans. Introducing this temporal aspect of the user input, a machine could make decisions based on, for instance, the user's mood, energy level, physical pain, integrity, or, based on experience, the user's most probable plan of action. To name a few examples, your house could sense that you are in a very good mood and that it is Friday evening. Based on the knowledge that your house has gathered over time about your character, it decides to play up-tempo music at a decent volume and set the lighting bright and colourful. Or your house could recognise the pattern of actions you typically perform when you are leaving, and, because it is morning, will start closing the windows and will turn off the heating. In another application area, a hospital could be equipped with webcams on each patient's bed, allowing for continuous monitoring of the pain that patients experience in terms

of its intensity and the type of pain (e.g. acute or prolonged).

Probably the most important aspect of human sensing is the understanding of a person's facial expressions. Facial expressions help to synchronise turn-taking, to signal comprehension or disagreement and to let a dialogue run more smoothly, with fewer interruptions. With facial expressions we clarify what is said by means of lip-reading, we stress the importance of the spoken message by means of conversational signals like raising eyebrows, and we signal boredom, emotions and intentions [Russell and Fernandez-Dols, 1997].

A major limitation of the majority of existing facial expression recognition systems is that they focus on emotion recognition only. However, displaying emotions is just one of the many functions of facial expressions. As stated above, the face sends many non-emotional communicative signals. Therefore we aim in this work to recognise instead every possible facial action by following the theory of the Facial Action Coding System (FACS, [Ekman et al., 2002]). This coding system defines atomic facial muscle actions called Action Units (AUs). With FACS, every possible facial expression can be described as a specific combination of AUs, including, but not limited to, expressions of emotions. A detailed introduction to FACS will be provided in section 2.1.4 of this thesis.

Existing AU recognition systems only focus on the morphological aspects of a facial expression. That is, they only report on what facial muscles, or AUs, were activated. The body of research in cognitive sciences which suggests that the temporal dynamics of human facial behaviour (e.g. the timing and duration of facial actions) are a critical factor for interpretation of the observed behaviour, is large and growing [Cohn and Schmidt, 2004, Bassili, 1978, Hess and Kleck, 1990]. Some researchers do use aspects of facial expression temporal dynamics to increase the recognition rates of facial expression recognition systems (e.g. [Littlewort et al., 2004, Kaliouby and Robinson, 2004]). However, despite their reported significance, no research group has yet endeavoured to analyse these facial expression temporal dynamics explicitly. In this thesis we will propose to do so for the first time.

Temporal dynamics of facial expressions contain an enormous amount of information about human behaviour. Facial expression temporal dynamics are essential for categorisation of complex psychological states like various types of pain and mood [de C. Williams, 2002]. They are also the key parameter in differentiating between posed and spontaneous facial expressions [Ekman and Rosenberg, 2005, Hess and Kleck, 1990]. For instance, it has been shown that spontaneous smiles, in contrast to posed smiles (like a polite smile), are slow in onset, can have multiple peaks of AU12 (rises of the mouth

corners, as in smiles) , and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1 second [Cohn and Schmidt, 2004].

One can discern two approaches to facial expression analysis: geometry-based approaches and appearance based approaches. Appearance based approaches use photometry-related properties of images such as colour, intensity, texture, or the presence of corners and edges. This includes approaches that are more abstract, but that still rely on the basic image intensity/colour properties of an image such as Gabor wavelet features or basic Principal Component Analysis (PCA) applied to the intensity values of an image.

Approaches using geometric features were proposed from the beginning of automatic facial expression analysis [Parke, 1974]. An approach using geometric features, in contrast to an appearance based approach, employs the geometrical properties of a face such as the positions of facial points relative to each other, the distances between pairs of points or the velocities of separate facial points. Recently it has been claimed that appearance based approaches are superior to geometry based approaches for the recognition of AU activation [Bartlett et al., 1999]. However, it has been shown that this is not always the case [Pantic and Patras, 2006]. In this thesis we will show that again.

In this thesis we have chosen to use geometric features. Bassili et al. [Bassili, 1978] already have shown that the movements of facial feature points are good descriptors of facial expressions, and based on these facial point movements it is theoretically possible to recognise 26 out of the 32 AUs [Bassili, 1978]. We believe that a geometric feature based approach performs as well as appearance based techniques for the recognition of AU activation. In situations of large head motion geometric feature based approaches may even outperform appearance based techniques, as the latter are very sensitive to registration errors. But more importantly, they provide a more intuitive approach to the analysis of facial action temporal dynamics. We will argue that with a geometry-based approach, extracting temporal dynamics attributes is intuitive and straightforward. There is a direct relationship between the velocity vectors of characteristic facial points on the one hand and the temporal dynamics of facial expressions on the other hand. Although it is not impossible to analyse facial temporal dynamics with appearance based approaches, a similar direct relationship does not exist.

The goal of this thesis is to present our research on the automatic analysis of the morphologic and temporal dynamic aspects of facial expressions. We will begin by giving an overview of the various theories that exist on ways to describe facial expressions (chapter 2) and an overview of existing

databases for facial expression recognition (chapter 3). We will then present our geometric feature based system.

The description of the system is split into four chapters. In chapter 4 we will describe how we localise 20 fiducial facial points in the first frame of a video. Then, in chapter 5 we will describe how we track these facial points and extract features from the tracking data. Finally, in chapter 6 we give a detailed description and evaluation of our methods to detect the activation of AUs in a video sequence and in chapter 7 we explain how we recognise the temporal patterns of AUs.

Three studies showing how our facial action analysis system can be used for human behaviour analysis are presented in chapters 8 and 9. While the main goal of this thesis is the automatic recognition and analysis of AUs, we have also investigated the possibility of recognising emotions. In chapter 8 we propose to use our geometric feature based approach to recognise emotions, either directly from the geometric features or in a two-step way in which we first detect AU activations and base our emotion recognition on the AU detection results. In chapter 9 we present two studies explaining how we have used our method to distinguish posed from spontaneous brow actions and to distinguish posed from spontaneous smiles. In the latter, not only have we used information from the face, but we have fused this data with cues from head and shoulder movements as well. Finally, in chapter 10 we provide concluding remarks about our work and give some directions for future research.

1.1 Contributions

This thesis offers the following innovative contributions to the field of *facial action analysis*:

1. This thesis presents a geometric feature-based method for fully automatic facial Action Unit analysis from near-frontal-view videos, integrating face detection, facial point detection, tracking and facial action analysis. The method attains state-of-the-art recognition results and can detect 22 out of 32 Action Units.
2. This thesis presents a method to fully automatically recognise the four temporal phases neutral, onset, apex, and offset of facial muscle actions in near-frontal-view videos for the first time in the literature.
3. This thesis introduces the MMI-facial-expression-database, which is the first searchable database publicly online available to contain videos of spontaneous facial expressions which are FACS

annotated, videos of synchronised frontal/profile view recordings of the face and a very large number of videos showing single-AU activations.

4. This thesis reports on using multiple affective cues, namely facial expression, head motion and shoulder actions, to distinguish between posed and spontaneous expressions fusing multiple visual cues for the first time in the literature.
5. This thesis provides insights into the ongoing question about the way emotions could be recognised by computers, by implementing different human cognitive models.
6. This thesis report on the award-winning live demonstration that integrates our facial action analysis systems with an artificial intelligent portrait painter. The system paints a sitter's portrait in an emotionally enhanced way using the image where the sitter showed maximum emotion.

The facial point detection and point tracking methods were developed by colleagues during the author's position as a PhD candidate in Dr. Maja Pantic's laboratory. The head tracker used in chapter 9 was developed by the robotics institute at Carnegie Mellon University, Pittsburgh USA, and the face detector used in chapter 4 was developed by Dr. Marian Bartlett's group at the University of California San Diego, USA. All other methods, as well as the integration of all the presented methods into one system, were developed by the author.

1.2 Application areas

A system that could enable fast and robust facial expression recognition would have many uses in both research and application areas as diverse as behavioural science, education, entertainment, medicine, and security. Following are some applications that we think are viable in the near future. Note that this list is just a very small slice of the possible applications.

- *Continuous pain monitoring of patients.* Monitoring the pain level of patients in hospital is a complicated task. The fact that it is primarily a mental state makes it hard to measure pain objectively. Currently nurses visit a patient five times a day and ask them to self-report their pain on a 5-point scale.

There are three major problems with this approach: first of all, with only five checks a day, it is discrete in time. Secondly it is an extremely subjective score, both from the point of view of the person describing the pain and the point of view of the person recording the pain. A patient might act tough and under-report his pain or might whine and exaggerate the pain felt. On the other hand, the nurse might *think* that the patient is exaggerating while this does not necessarily have to be the case. The third major issue is that the method used for keeping this pain diary does not differentiate between different types of pain (e.g. acute or prolonged pains).

An automatic facial expression recognition system could solve all three problems in one go. It has been shown that it is possible to derive a measure of pain and to distinguish between different types of pain from a patients' facial expressions [de C. Williams, 2002]. Finally, if such a system would be able to run in at least near-real time, an incredible increase in accuracy and reliability of measurements as well as an increased reaction speed to acute pain could be achieved, compared with the current five-times a day measurement.

- *Avatars with expressions.* Virtual environments have become tremendously popular in the 21st century. In gaming, the immensely popular massive multiplayer roleplaying game World of Warcraft (WoW) has 8.5 million users, while Secondlife is a virtual world with over 1.5 million active users. In Secondlife, interaction with other users and enhancing the appearance of your avatar are considered very important. Secondlife has an entire economy of its own, with people spending about 2 million US Dollars every day on new clothing, cars, houses etc. If the avatars were able to mimic their user's facial expressions recorded by a webcam and analysed by a facial expression recognition system, the level of immersion in the virtual world would increase. It would make interaction with other users more realistic and possibly easier. The automatic analysis of facial expressions on the client-side of Secondlife would ensure minimal use of bandwidth, as only very little data has to be send to describe a facial expression in terms of its Action Units.

Apart from improved avatar-avatar interaction (and thus, through this layer of abstraction, human-human interaction), the virtual environment and its agents could benefit from reading the user's expressions. It would open up possibilities for a more varied dialogue between an artificial intelligent agent and a user's avatar, where the agent reacts appropriately to happy or angry faces. This would significantly increase the level of immersion. Also, if the user is immersed in a gaming environment, the game could adapt its difficulty level based on information from the facial expressions of the user. For example, if the user is playing a shoot-em-up game and he is

looking bored, the game could decide to send more enemies, or increase the tactical intelligence of the existing enemies.

- *Smart homes.* Remote facial expression recognition would be an important element of the human sensing capabilities of ambient intelligent homes of the future. By measuring the facial expressions of the inhabitants over time, the house could adapt atmospherical parameters such as lighting and music. Facial expressions could also be used as instant commands. This would be especially efficient in combination with gaze direction recognition. For instance, a frown made while looking at the stereo could make it skip this song and start playing the next one.
- *Affective robots.* Traditional robots operate in environments that do not contain humans, mainly in factories or environments that are inherently hostile to humans (e.g., near bombs or in nuclear power plants). However, there has been a major shift towards so-called social robots. In contrast, these robots are designed to have contact with humans. This new generation of robots has been designed for tasks such as helping the elderly and the disabled in their homes [Forlizzi, 2005] or as robot pets and toys.

Because the whole idea behind these robots is communication with humans, it is very important to improve the current state of human-robot communication and interaction so that it will run smoother and with fewer mistakes. The Social Robots Project at Carnegie Mellon University states its mission as ‘wanting robots to behave more like people, so that people do not have to behave like robots when they interact with them’¹. To attain such human-robot interaction, it is of paramount importance for the robot to understand the human’s facial expressions.

- *Detection and treatment of depression and anxiety.* According to recent studies, anxiety disorders cost the United States of America 46.6 billion USD in 1990, nearly one-third of the nation’s total mental health bill of 147 billion USD. The USA National Institute of Mental Health reported that 90% of the people with emotional illness will improve or recover entirely if they get treatment². Research based on the FACS has shown that facial expressions can predict the onset and remission of depression, schizophrenia, and other psychopathological afflictions [Ekman and Rosenberg, 2005], can discriminate suicidally from non-suicidally depressed patients [Heller and Haynal, 1994] and can predict transient myocardial ischemia in coronary patients [Rosenberg et al., 2001]. FACS has also been able to identify patterns of

¹CMU social robots project, <http://www.cs.cmu.edu/social>. Date of access: June 29, 2007

²Health discovery, <http://health.discovery.com/centres/mental/articles/signofanxiety/signofanxiety.html>, Date of access: June 29 2007

facial activity involved in alcohol intoxication that observers not trained in FACS failed to note [Sayette et al., 1992]. This suggests there are many applications for an automatic facial expression recognition system based on FACS.

1.3 Publications

1. M.F. Valstar, I. Patras and M. Pantic, “Facial action unit recognition using temporal templates”, *Proc. IEEE Int’l Workshop on Robot Human Interaction (ROMAN)*, Kurashiki, September 2004
2. M.F. Valstar, I. Patras and M. Pantic, “Motion history for facial action detection in video”, *Proc. IEEE Int’l Conf. Systems Man and Cybernetics*, vol 1, pp. 635-640, Den Haag, October 2004
3. M.F. Valstar, I. Patras and M. Pantic, “Facial Action Unit Detection using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data”, *IEEE Int’l Conf. Computer Vision and Pattern Recognition*, vol. 3, pp. 76-84, San Diego, June 2005
4. M. Pantic, M.F. Valstar, R. Rademaker and L. Maat, “Web-based database for facial expression analysis”, *Proc. IEEE Int’l Conf. on Multimedia and Expo (ICME ’05)*, pp. 317-321, Amsterdam, The Netherlands, July 2005
5. M. Pantic, I. Patras and M.F. Valstar, “Learning spatiotemporal models of facial expressions”, *Proc. Int’l Conf. Measuring Behaviour (MB’05)*, pp. 7-10, Wageningen, The Netherlands, September 2005
6. M.F. Valstar and M. Pantic, “Fully automatic facial action unit detection and temporal analysis”, *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition (CVPR’06)*, New York, USA, June 2006
7. M.F. Valstar and M. Pantic, “Biologically vs. logic inspired encoding of facial actions and emotions in video”, *Proc. IEEE Int’l Conf. on Multimedia and Expo (ICME ’06)*, Toronto, Canada, July 2006
8. M.F. Valstar, M. Pantic, Z. Ambadar and J.F. Cohn, “Spontaneous vs. posed facial behavior: automatic analysis of brow actions”, *Proc. ACM Intl. conf. on Multimodal Interfaces (ICMI’06)*, Banff, Canada, November 2006

9. M.F. Valstar and M. Pantic, “Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics”, *Proc. IEEE W’shop on Human Computer Interaction (HCI’07), in conjunction with IEEE ICCV’07*, Rio de Janeiro, Brasil, October 2007
Best Paper Award
10. M.F. Valstar, H. Gunes and M. Pantic, “How to Distinguish Posed from Spontaneous Smiles using Geometric Features”, *Proc. ACM Intl. conf. on Multimodal Interfaces (ICMI’07)*, Nagoya, Japan, November 2007
11. S. Colton, M.F. Valstar and M. Pantic, “Emotionally Aware Automated Portrait Painting”, *Proc. Int’l Conf. Digital Interactive Media in Entertainment & Arts (DIMEA’08)*, Athens, Greece, September 2008

Chapter 2

Facial Expression Analysis

The face is an extremely familiar yet fascinatingly surprising thing. We are accustomed to see faces many times a day, yet the same faces that are most familiar to us also have the ability to riddle us time and time again. Although we can invoke clearly in our mind the image of a characteristic smile or frown of someone we know well, many times we find ourselves thinking “What does she really want?” when an unexpected expression is displayed by the same person. They say that some people’s faces can be ‘read like a book’, while others have a ‘poker face’ that apparently reveals nothing of their feelings or intentions. It is this paradox of the obvious and the mysterious that intrigues people to study the face.

The face is our main means of interaction with the world around us. It harbours all five senses and is the exclusive site of four of them: smell, taste, vision and hearing. The face is our means to identify people, to assess age, ethnic group, beauty, and the sex of an individual. It is our most important mode of non-verbal communication [Ambady and Rosenthal, 1992], with a wealth of communicative signals coming from the facial expressions, head movements and gaze directions [Ekman and Friesen, 1969]. It is therefore not strange that humans have been studying facial expressions since ancient times.

The first studies of facial expressions belonged to the realm of philosophy, with great thinkers like Aristotle and Stewart theorising about the form and function of emotions. Greek drama masks associated the appearance with both emotion and character, while the first treatise on the face and facial expressions is from ca. 340 B.C. and stated that physiognomic traits were best discerned by comparing people with animals [Fridlund, 1994]. The author, unknown, but thought to be Aristotle, did not regard the reading of emotions as particularly important.

With Darwin and his contemporary Duchene, the study of facial expressions became an empirical study. Darwin studied the similarities and differences of displays of emotion between humans and animals in his book *The Expression of the Emotions in Man and Animals* [Darwin, 1872], concluding that “the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements”. Duchene [de Bologne, 1862] used electrical stimuli on faces of living and dead prisoners to determine which facial muscles belonged to each facial expression.

Darwin’s studies created large interest among psychologists and cognitive scientists. The 20th century saw many studies relating facial expression to emotion and inter-human communication. Most notably, Paul Ekman reinvestigated Darwin’s work and claimed that there are six universal emotions, which are produced and recognised independently of cultural background [Ekman and Rosenberg, 2005]. However, not all scientists agree with this. People like Alan Fridlund argued that emotions do not mirror what we feel, instead they are used to coerce other people into a desired action [Fridlund, 1994]. A smile might be used to persuade someone to proceed or come closer while an angry face might signal that the second party should back off, or face the consequences.

Since 1974 [Parke, 1974] the investigation of facial expressions was joined by the computer science community which, inspired by the findings of the cognitive scientists, endeavoured to automate facial expression analysis from videos or still images. Many different approaches were used, ranging from holistic methods based on colour or intensity, to feature based approaches that use the positions of a number of facial points to classify facial expressions.

In the following I will give an overview of the related work in automatic facial expression analysis. Because, as stated above, automatic facial expression analysis is often inspired by and builds upon the research by cognitive scientists, I will start in section 2.1 with an overview of the most important theories in that field, including various proposals of coding systems to classify facial expressions. We will argue that the Facial Action Coding System, discussed in section 2.1.4 is the preferred coding system. Next, in section 2.2 we will discuss the the most important works in facial expression recognition, keeping in mind our choice of FACS as our coding system.

2.1 Facial affect theory and facial action coding

When we start designing a system that is to classify facial expressions, probably the first question that jumps to mind is ‘What are the atomic elements of a facial expression’, i.e. what are the classes our classifiers have to learn? Let us present first the relevant findings of cognitive scientists.

There are two different ways to analyse facial expressions: one considers facial affect (emotion) and the other facial muscle actions [Pantic and Rothkrantz, 2000, Pantic and Rothkrantz, 2003, Tian et al., 2005]. These two streams stem directly from the two major approaches to facial expression measurement in psychological research [Cohn, 2006]: message and sign judgement. The aim of the former is to infer what underlies a displayed facial expression, such as affect or personality, while the aim of the latter is to describe the *physical appearance* of the shown behaviour, such as facial movement or facial component shape.

While message judgement is all about interpretation, sign judgement is agnostic, independent from any interpretation attempt, leaving the inference about the conveyed message to higher order decision making. The most commonly used facial expression descriptors in message judgement approaches are the six basic emotions (anger, disgust, fear, happiness, sadness and surprise) [Cohn, 2006] proposed by Ekman [Ekman and Friesen, 1969, Ekman, 1992], Plutchik’s emotion wheel [Plutchik, 1980] and Russell’s circumplex model [Russell, 1980]. These coding systems all address fairly basic emotions. Baron-Cohen [Baron-Cohen et al., 2004] proposed a taxonomy of more complex mental states, describing the message of a facial expression in terms as ‘confused’, ‘concentrating’, or ‘attentive’. The most popular sign judgement systems are the Facial Action Coding System (FACS) and MPEG-4. The last two systems, FACS and MPEG-4, have many elements in common. Yet FACS is used to describe facial muscle actions and is used both by computer scientists and psychologists while MPEG-4 codes the actual movement of facial components and is used exclusively by computer scientists. In the following subsections these various coding systems will be explained.

2.1.1 Ekman’s Six Basic Emotions

The most commonly used facial expression coding system in the message judgement approach relates to the six basic emotions proposed by Ekman [Ekman and Friesen, 1969]. He suggests that the *six basic emotions* anger, disgust, fear, happiness, sadness and surprise have evolved in the same way for

all mankind. Moreover, the theory suggests that our facial expressions show these emotions and that these displays of emotion are produced and recognised universally, that is, independent of culture or other nurture effects [Keltner and Ekman, 2004]. These six basic emotions are thus linked to a small number of universal prototypic facial expressions.

This theory of facial expressions mirroring a fixed set of universal emotions was inspired directly by Darwin's observations that both humans and animals display similar expressions in similar situations [Darwin, 1872]. Both dogs, cats and humans displayed a particular expression when they were in pain, felt affection or had great anxiety. The way the expression was shown was different for each race, but the underlying so-called 'felt' emotion that caused the expression was the same. As we do not share a culture with the primates he studied, his observations led Darwin to the conclusion that our facial expressions belong to our evolutionary, rather than to our cultural, heritage. These observations, published in 1872 were left undebated for almost 100 years. Then, in the 1960s, Paul Ekman picked up his research in an effort to validate Darwin's theories.

Ekman studied a culturally isolated New Guinean society (the South Fore), first from film and later in person [Ekman, 1982]. From the films he noted that he 'saw nothing [he] hadn't seen before'. When he visited the stone-age society himself he subjected the South Fore people to an experiment of emotion recognition. He showed them pictures of people from different societies displaying an expression of affect. According to his study, the tribesmen were capable of recognising six basic emotions, thereby showing that these emotions were hardwired in our genetics instead of learned from our culture.

Dacher Keltner, who studied emotions together with Paul Ekman, said that "Our facial expressions of emotions are the products of evolution, to help us survive, reproduce, and get along". While this might be true, it does not guarantee that what we feel is directly and automatically shown in our facial behaviour. Humans are great liars [Smith, 1995] and to be able to live in a social group, much evolutionary effort seems to have been put into creating techniques to disguise our true feelings.

So if we both show a facial expression when we truly feel an emotion as well as when we don't feel the emotion but only want to display the expression of that emotion for some social effect, why would we use the shown facial expression as an indicator of the felt emotion? Ekman acknowledges this problem but in later studies also showed that it is possible to recognise from barely detectable facial cues whether a shown expression of affect is genuine or not. According to Ekman, differences in the temporal dynamics of facial expressions and of co-occurrences of certain facial muscle actions can

indicate whether an expression was genuine (we explore these claims rigorously in chapter 9). Even if this is true¹ and we can either recognise the felt emotion or recognise that the displayed expression does not reflect the subject’s felt emotion, we do not think Ekman’s system is suitable to describe facial expressions.

Most importantly, while Ekman concluded that these six basic emotions are universally recognised and produced, he never claimed that these are the only possible emotions nor the only or fundamental facial expressions. And his system does not allow us to describe the other emotions; with Ekman’s system of six basic emotions, we cannot describe social signalling such as a brow flash, which is often used in western society to greet a known person, or brow lowering to indicate that one doesn’t understand or agree with the other person. So, while there might be a link between facial expression and emotion, the six basic emotions do not allow us to describe all possible facial expressions and is thus not suitable for an automated facial expression analysis system. It is a judgement system that does not attempt to encode all possible meanings of a facial expression, only a small number of universal emotions. Our goal is to be able to analyse all possible facial expressions and this coding scheme is therefore not suitable for our system.

2.1.2 Plutchik’s Emotion Wheel

While Ekman proposed that there are six basic emotions, Plutchik [Plutchik, 1980] on the other hand, proposed eight primary emotions: acceptance, anger, anticipation, disgust, fear, joy, sadness and surprise; every other emotion can be produced by mixing the primary ones. To this aim he introduced the *emotion wheel*, which is similar to a palette of colours. The three dimensional circumplex model describes the relationships between concepts of emotion with each emotion smoothly flowing into another when we traverse a path on the surface of the 3-dimensional cone shown in Fig. 2.1.

The cone’s vertical dimension represents the intensity of the emotion and the position on the circle defines in what basic emotion sector we are. The eight sectors are designed to indicate that there are eight primary emotion dimensions, formed by four pairs of opposites. Secondary emotions are produced by combinations of primary emotions that are adjacent on the emotion wheel. For instances, Plutchik characterises “love” as a combination of “joy” and “acceptance”, whereas “submission” is a combination of “acceptance” and “fear”.

¹We have reason to believe this it is possible to detect whether an expression was posed or spontaneous, see section 9

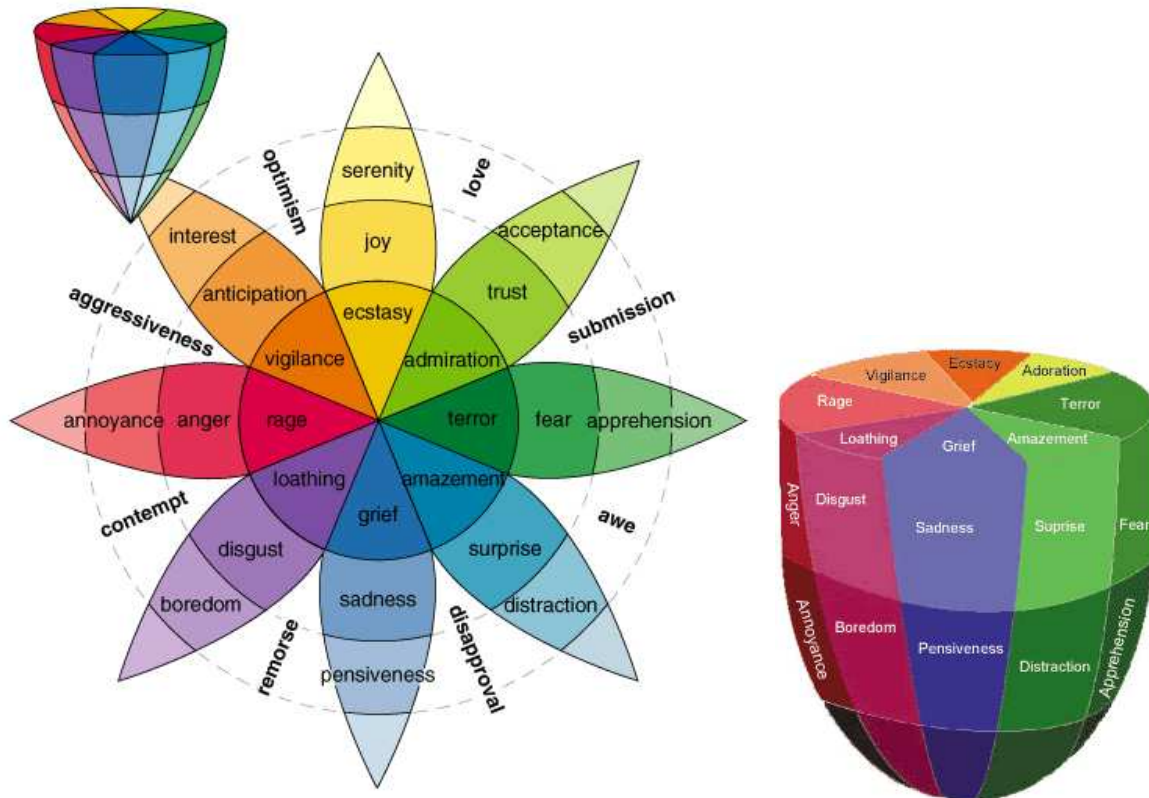


Figure 2.1: Plutchik's emotion wheel (left) and the accompanying emotion cone (right).

The reason for inclusion of the eight primary emotions is that they all have a direct relation to adaptive biological processes. For instance, when we gain a valued object, the cognitive process is that we 'possess' and thus we feel 'joy'. The associated action would be to retain this state or repeat the process, gaining more valued objects and thereby feeling joy again. The theory is very much based on our inner state, i.e. on the way we actually feel.

This theory has similarities with Fridlund's theory which states that our shown expressions are learned, socially influenced tools to obtain a goal. However, the same argument that we made against the six basic emotions holds here: nothing guarantees that what we feel is automatically shown by our facial expressions. Being a judgement system that is based on feeling, it is problematic to use this system to describe non-emotional communicative signals such as a brow-flash used in greetings. Therefore, while Plutchik's emotion wheel might be a good classification system of what we feel, it need not be a good classification system to describe the shown facial expression.

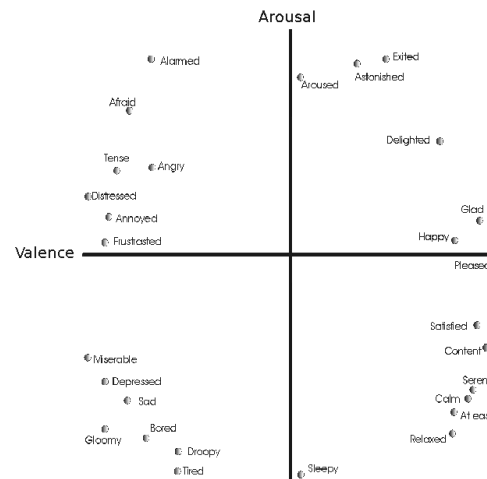


Figure 2.2: Russell's circumplex model of emotions, showing the emotion space spanned by the bipolar continui *valence* and *arousal*.

2.1.3 Russell's Circumplex model

Many cognitive scientists oppose the theory of a set of discrete, basic emotions [Mandler, 1984, Ortony and Turner, 1990, Russell, 1995, Turner and Ortony, 1992]. Some of these opponents instead take a dimensional view of the problem. In their view, affective states are not discrete and independent of each other, instead they are systematically related to one another [Cacioppo and Berntson, 1994, Mehrabian and Russell, 1973, Smith and Ellsworth, 1985].

Perhaps the most influential of the scientists who proposed this systematic view is James A. Russell. He proposed a system of two bipolar continui, namely *valence* and *arousal* [Russell, 1980]. Valence roughly ranges from sadness to happiness while arousal ranges from boredom or sleepiness to frantic excitement. According to Russell, all emotions lie on a circle in this two dimensional space (see Figure 2.2).

These dimensions are commonly treated as independent factors, but real-world experience suggests that they are often correlated. Negative stimuli tend to have a higher intensity (few pleasant things are felt as intensely as truly unpleasant things), and higher intensity tends to amplify valence (a very large steak is much more pleasant than a small steak). Interactions of this nature often make it difficult to dissociate the 2 dimensions. This results in self-reported emotions not lying on a circle, as theorised by Russell, but more in V-like shapes [Bradley and Lang, 2000, Merckx et al., 2007].

While a continuous space could possibly represent all possible facial expressions, this is not guaranteed



Figure 2.3: Some examples of Action Units. Action Units are atomic facial muscle actions described in FACS.

by Russell’s theory and indeed has not been shown. It is unclear how a facial expression should be mapped to the space or, vice versa, how to define regions in the valence/arousal space that correspond to a certain facial expression. Being a judgement system that is based on feeling, it is again problematic to use this system to describe non-emotional communicative signals. All this leads us to the conclusion that this system is not to be preferred for our purposes.

2.1.4 Facial Action Coding System (FACS)

The Facial Action Coding System (FACS) [Ekman et al., 2002] is the most widely used sign judgement system. The FACS associates facial expression changes with the actions of the muscles that produce them. It defines 32 atomic actions: 9 action units (AUs) in the upper face, 18 in the lower face, and 5 AUs that cannot be classified as belonging to either the upper or the lower face (for examples, see figure 2.3). Besides the AUs, there are also so-called action descriptors (ADs). There are 11 action descriptors for head position, 9 for eye position, and 14 for miscellaneous actions. Tables 2.2, 2.3 and 2.4 lists all 32 AUs.

AUs are considered to be the smallest visually discernable facial movements. They are atomic, meaning that no AU can be split into two or more smaller components. Any facial expression can be uniquely described by a combination of AUs. The FACS also provides the rules for recognition of AUs’ temporal segments (onset, apex and offset) and provides guidelines on how to score the intensity of a facial action on a five-point scale.

Using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into the specific AUs and their temporal segments that produced the expression. As

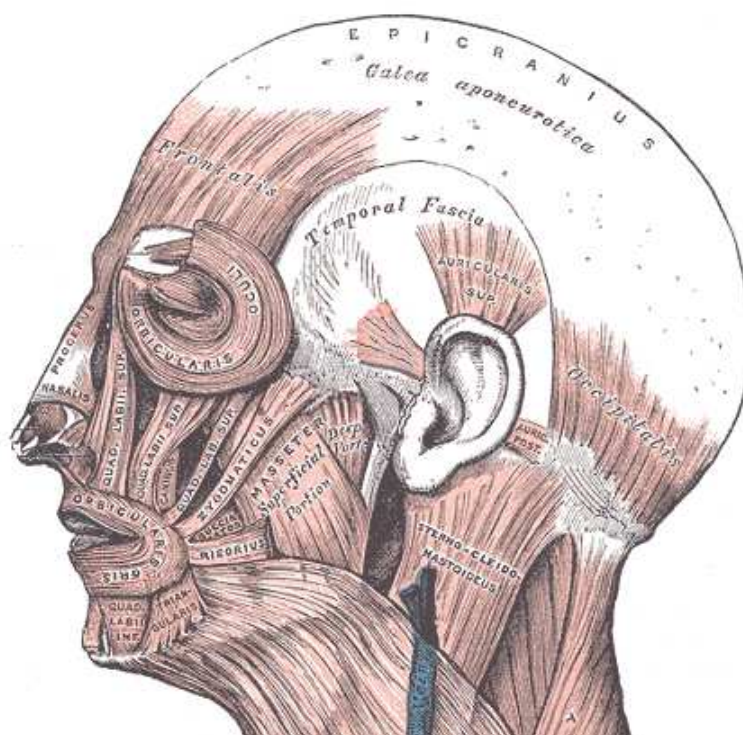


Figure 2.4: The facial muscles seen from the side (image taken from Gray's Anatomy)

AUs are independent of any interpretation, they can be used by any higher order decision making process including recognition of basic emotions [Ekman et al., 2002], or pre-programmed commands for an ambient intelligent environment.

As stated above, the FACS has an anatomical basis. Facial muscles were systematically studied for the first time by Duchene [de Bologne, 1862]. Since then many studies have been carried out on the functioning of our facial muscles. Figure 2.4 shows the major facial muscles. There are about fifteen muscles that are primarily used to create facial expressions. These muscles are described in greater detail in Faigan's work [Faigan, 1990]. The muscles of the upper half of the face frequently used in facial expressions are illustrated in figure 2.5, and the muscles of the lower half of the face in figure 2.6. The muscle used to produce AU5 (raising of the upper eyelid) can't be seen from these images, as it actually resides on the inside of the eye socket. Figure 2.7 illustrates how the elevator palpebrae superior is attached to the eyelid. A brief description of these muscles is presented here. We use the terminology of the FACS; where the muscle *emerges* is where it is attached to the bone, and where the muscle *attaches* is where it connects with the soft tissue of the face.

1. *Orbicularis oculi*: This muscle, circumfering the eyes, emerges from the inner side of the eye socket and attaches to the skin of the cheek and the eyelids. It is used for squinting, narrowing

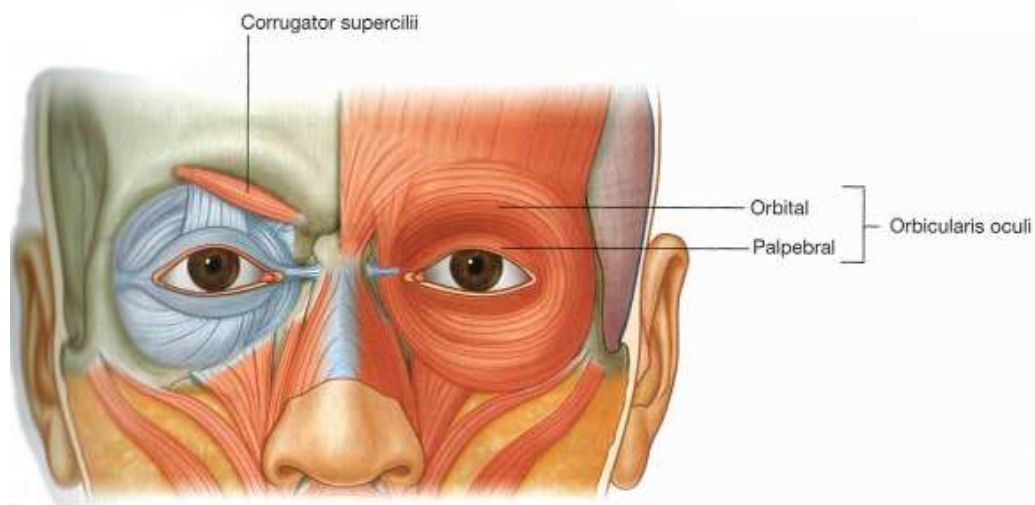


Figure 2.5: Upper face muscles used to produce facial expressions (image taken from Gray's Anatomy).

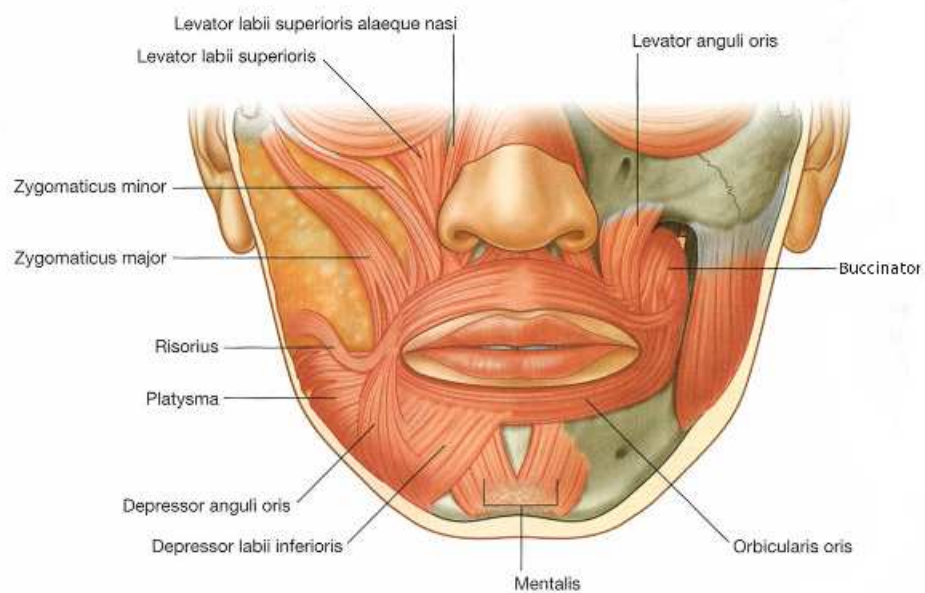


Figure 2.6: Upper face muscles used to produce facial expressions (image taken from Gray's Anatomy).

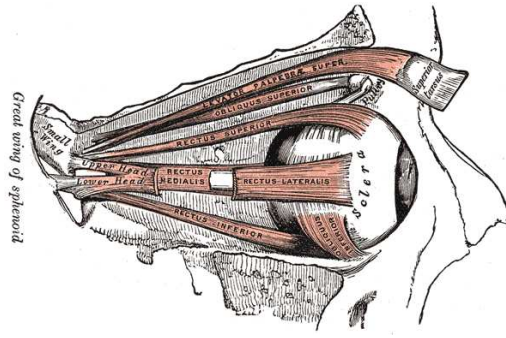


Figure 2.7: Illustration of the elevator palpebrae superior, used to raise the upper eyelids (AU5, image taken from Gray's Anatomy).

of the eye, and raising the cheeks.

2. *Levator Palpebrae*: This muscle emerges from inside the eyesocket, above the eye, and attaches to the upper eyelid. It is used to open the eyes. As such, it is normally contracted a little when we have the eyes open. Stronger contractions occur when we display, for example, surprise.
3. *Levator Labii Superioris*: Emerges from the medial infra-orbital margin, i.e. the lower part of the cheekbone, and is attached to the skin and tissue of the upper lip. Point of attachment is halfway between the mouth corner and the philtrum². This muscle is used in sneers or displays of disgust.
4. *Zygomaticus major*: This muscle emerges from the outside of the zygomatic bone, i.e. the cheekbone. It attaches to the skin and tissue of the upper lip, just inward of our mouth corners. This muscle produces smiles.
5. *Zygomaticus minor*: This muscle runs parallel to the zygomaticus major, but more inward, towards the nose. As such, it attaches to the upper lip a bit more inward too, halfway between the Zygomaticus major and the levator labii superioris. It emerges from the lateral infra-orbital margin, above and inward relative to where the zygomaticus major emerges.
6. *Risorius*: Emerges from the deep fascia of the side of the face and the parotid region which extends to the neck. It attaches to the mouth corner and is used to widen the mouth laterally, such as in displays of fear.
7. *Frontalis*: This muscle originates near the top of the skull and attaches to the skin and tissue under the eyebrows. It is a very wide muscle, and not all parts of it need be contracted at the

²The philtrum is the vertical depression in the centre of the upper lip directly under the tip of the nose

same time. As such, it is responsible for raising both the inner and the outer eyebrows.

8. *Orbicularis oris*: This muscle, circumfering the mouth, emerges near the midline on the anterior surface of the maxilla and the mandible. It attaches to the mucous membrane of the margin of the lips and the tissue directly above and under the lips.
9. *Corrugator Supercilii*: This muscle emerges from the medial superciliary arch, runs across the nasal bridge and attaches to the skin of the medial forehead, under the inner portion of the eyebrows. It is used to lower the inner eyebrows and wrinkles the skin between the eyebrows. Four relatively distinct movements can be produced by orbicularis oris, a pressing together, a tightening and thinning, a rolling inwards between the teeth, and a thrusting outwards of the lips.
10. *Procerus*: This muscle emerges from the nasal bone and attaches to the skin of the medial forehead. It is used to wrinkle and frown the forehead.
11. *Depressor anguli oris*: Also called triangularis, this muscle emerges from the lower margin of the sides of the jaw and attaches to the tissue and the skin under the mouth corner. It is used to lower the mouth corners.
12. *Depressor labii inferioris*: This muscle emerges from the lower margin of the sides of the jaw and attaches to the skin of the lower lip. It is used to depress the lower lip laterally.
13. *Mentalis*: This muscle emerges from middle of the lower jaw and attaches to the skin of the chin. It is used to elevate and wrinkle the skin of the chin and protrude the lower lip.
14. *Nasalis*: This muscle actually consists of three muscles: the compressor, dilator and depressor. They emerge from the frontal process of maxilla and attach to the nostrils. Their function is opening (flaring) and closing the nostrils, e.g. in forced respiration.
15. *Levator anguli oris*: This muscle emerges from the anterior surface of the maxilla below the infraorbital foramen (between the cheekbone and the nose) and attaches to the outer end of the upper lip. It elevates the angle of the mouth and as such is used in smiles.

A point of caution is justified concerning the list of muscles given above. Although the list above summarises the most important muscles involved in the activation of AUs, the list is by no means exhaustive. To give an indication of the complexity of the problem, some human anatomy researchers

AU	Muscle	AU	Muscle
1	Frontalis (medial)	2	Frontalis (lateral)
4	Depressor supercillii, Corrugator, Procerus	5	Levator Palpebrae
6	Orbicularis oculi (pars orbitalis)	7	Orbicularis oculi (pars palpebralis)
8	Orbicularis oris	9	Levator Labii Superioris (alaeque nasi)
10	Levator Labii Superioris (caput infraorbitalis)	11	Zygomatic minor
12	Zygomaticus major, Levator anguli oris	13	Caninus, Levator anguli oris
14	Buccinator	15	Depressor anguli oris, Triangularis
16	Depressor labii inferioris	17	Mentalis
18	Incisivus Labii	20	Risorius
21		22	Orbicularis oris (outer part), Buccinator
23	Orbicularis oris (inner part)	24	Orbicularis oris (inner part)
25	Depressor labii inferioris, Mentalis, Orbicularis oris	26	Temporalis, Masseter
27	Digastric, Temporalis, Masseter	28	Orbicularis oris
31		38	Nasalis (pars alaris)
39	Nasalis(pars transversa), Depressor septi nasi	43	Orbicularis oculi
45	Levator palpebrae, Orbicularis oculi	46	Orbicularis oculi

Table 2.1: Facial muscles used in Action Units.

have mentioned that there are more than 15 muscles involved in producing the 32 AUs. One anatomy researcher noted that there are 36 named muscles of facial expression³. Sometimes there are more than one muscles involved in producing a single AU, and sometimes the same AU is involved to produce a number of different AUs. For example, in a genuine smile, there are typically six muscle types involved: the zygomaticus major and minor, the orbicularis oculi, the levator labii superioris, the levator anguli oris, and the risorius⁴. This however, would be coded only as AU12 + AU6 and depending on the strength of the smile, some of these muscles might not be used after all. Thus, although FACS is based on muscle activations, there is no one-to-one mapping of facial muscle actions to facial action units. This is not a problem however, as it still describes all the atomic visible facial actions.

Table 2.1 gives an overview of what muscles are involved in each AU.

³Muscles to smile, muscles to frown, <http://anatomynotes.blogspot.com/2006/01/muscles-to-smile-muscles-to-frown.html>, Date of access: July 27 2007

⁴Does it take fewer muscles to smile than it does to frown?, <http://www.straightdope.com/columns/040116.html>, Date of access: July 27, 2007




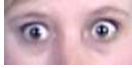




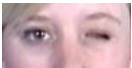
	AU1 Raised inner eyebrow		AU2 Raised outer eyebrow
	AU4 Eyebrows lowered and drawn together		AU5 Raised upper eyelid
	AU6 Raised cheek, compressed eyelids		AU7 Tightened eyelids
	AU43 Eyes closed		AU45 Blink
	AU46 Wink		

Table 2.2: Upper facial Action Units.

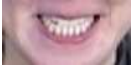
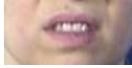





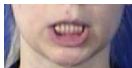
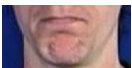

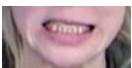
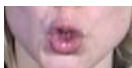




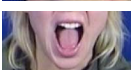

	AU8 Lips towards each other		AU10 Raised upper lip
	AU11 Deepened nasolabial furrow		AU12 Lip corners pulled up
	AU13 Lip corners pulled sharply up		AU14 Dimpler - mouth corners pulled inwards
	AU15 Lip corners depressed		AU16 Lower lip depressed
	AU17 Chin raised		AU18 Puckered lips
	AU20 Mouth stretched horizontally		AU22 Lip funneled and protruded
	AU23 Lips tightened		AU24 Lips pressed
	AU25 Lips parted		AU26 Jaw dropped
	AU27 Mouth stretched open		AU28 Lips sucked into the mouth

Table 2.3: Lower facial Action Units.

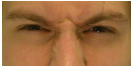




	AU9 Nose wrinkler		AU21 Neck tightened
	AU31 Jaw clenched		AU38 Nostril wings flared out (left is neutral, right active)
	AU39 Nostril wings compressed (left is neutral, right active)		

Table 2.4: Action Units belonging to neither the upper nor the lower facial area.

FACS 2nd edition vs. 1st edition

The theory for FACS was published in 1976 [Ekman and Friesen, 1976] by Paul Ekman and Wallace V. Friesen and the original FACS manual was published shortly after in 1978 [Ekman and Friesen, 1978]. While using the system for several years in their lab and training new FACS coders, they updated the rules and definitions of the system. At first the changes were handed out to the new FACS coders in the form of an addendum. However, as changes became more structural, a new version of FACS was needed.

In 2002, a new version of FACS [Ekman et al., 2002] was finally published, this time with a third author, Joseph Hager. Most co-occurrence rules were removed⁵, a number of AUs were made redundant, minimum requirements were eliminated, and a novel intensity scoring definition was introduced. Unfortunately, the authors decided not to rename the system. It is still simply known as FACS, not as ‘FACS2’, ‘FACS 2002 revision’ or ‘FACS version 2’. The lack of a means to distinguish the 1978 and the 2002 versions by nomenclature has led to confusion in the facial expression community. Some papers mention that there are 46 AUs, others mention 32. Some allow scoring of AU27 and AU25 occurring together (possible in the new version), others don’t. In this thesis, we will adhere to the 2002 version of FACS, and, where a distinction between the versions is needed, we will refer to them as FACS1 and FACS2. For clarity, the changes between FACS1 and FACS2 are listed below:

- The AUs AU41 (drooped eyelid), AU42 (slit eyes), and AU44 (squinted eyes) are re-assigned to intensities of AU43 (eyes closed).
- The minimal requirements (intensity threshold for scoring) have been removed.
- A five-point intensity scale for every AU has been introduced.
- Co-occurrence rules have been eliminated.

2.1.5 MPEG-4

MPEG-4 is a multimedia compression standard that is capable of encoding each of the objects in a scene separately. It has a special model for encoding faces and facial expressions. The neutral face is

⁵Co-occurrence rules indicated that the scoring of certain AUs would disallow scoring of certain other AUs, even if their associated muscles were contracted. An example is AU27 (jaw stretched vertically) negating the scoring of AU25 if the co-occurrence rules were to be used.

specified by a 3D face model defined by the face definition parameters (FDP) and the expression on the face is described by facial animation parameters (FAP). See [Tekalp and Ostermann, 2000] for a description of face modelling in MPEG-4.

The MPEG-4 standard leaves a lot of the definition details of the FDP and FAPs to the designers of FERS. The facial mesh of the neutral face is made up from 84 well-defined facial feature points and there are 66 low-level FAPs. However, the FAPs have simple geometric definitions, such as 'Vertical displacement of midpoint between left corner and middle of top inner lip' for FAP-8. But the FAP-definitions describe displacements in one dimension and the designers of the FERS should work out how a single multi-dimensional motion of facial points is translated to a combination of FAPs.

The designers of MPEG-4 claim that the FAPs are based on muscle activations. However, their relation to facial muscle activations is a lot looser than the relationship between FACS and muscle activations. A single muscle action will commonly trigger many FAPs. For instance, in a smile (AU12), both FAPs that describe the left and right corner of the mouth will be triggered, and for both mouth corners a separate FAP will describe horizontal and vertical motion. Another issue is that the MPEG-4 FAPs can only encode displacements of visible facial feature points. Facial expressions that only deform the shape of the skin without moving any facial feature points (e.g. dimplers or deepening the nasolabial furrow), cannot be encoded. So while FAPs are indeed *related* to the facial muscles that underly facial expression, we definitely don't have a *one-to-one* mapping between FAPs and facial muscle actions.

The MPEG-4 FDP/FAP are created with the purpose of driving virtual animated faces. Its encoding is tailored towards computer graphics designers, and as such it is too detailed for a FERS. This unnecessarily complex encoding combined with the inability to describe a number of facial actions which induce changes in facial texture rather than displacements of facial points, made us decide not to use this facial action coding system.

2.2 Automatic Facial Expression Analysis

Automatic facial expression analysis has been an active research area in the fields of computer vision and pattern recognition since 1978 [Suwa et al., 1978]. In 1992, Samal and Iyengard gave an overview of the early work in this field [Samal and Iyengard, 1992]. Since 1978, the focus of attention has shifted a couple of times. First attention shifted from the analysis of still images to analysis of face video.

Later some researchers moved away from detecting a small number of prototypic expressions, such as emotions, to the detection of atomic facial actions; either FACS AUs or MPEG-4 FAPs. Recently the focus has shifted to two new aspects of facial expression analysis: the analysis of spontaneously displayed expressions and the analysis of the temporal dynamics of facial expressions. Indeed, the proposed systems in this thesis were specifically designed to address the issue of facial expression temporal dynamics and the system has been applied to spontaneous facial expression data repeatedly (see chapter 9).

As noted above in section 2.1, the two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action detection [Pantic and Rothkrantz, 2000, Pantic and Rothkrantz, 2003, Tian et al., 2005]. These streams stem directly from the two major approaches to facial expression measurement in psychological research [Cohn, 2006]: message and sign judgement.

Most automatic facial expression analysis systems developed so far are to be considered message judgement systems. They target human facial affect analysis and attempt to recognise a small set of prototypical emotional facial expressions like happiness and anger [Pantic and Rothkrantz, 2000, Pantic and Rothkrantz, 2003, Tian et al., 2005]. Even though automatic recognition of the six basic emotions from face images and image sequences recorded under controlled conditions is considered largely solved, reports on novel approaches are still being published even to this date (e.g., [Anderson and McOwan, 2006, Chang et al., 2006, Guo and Dyer, 2005, Zhang and Ji, 2005]). Exceptions from this overall state of the art in machine analysis of human facial affect include few tentative efforts to detect cognitive and psychological states like interest [Kaliouby and Robinson, 2004], pain [Bartlett et al., 2006], and fatigue [Gu and Ji, 2005].

While message judgement studies distinguish only a very limited number of facial expressions, human experts can manually code nearly any anatomically possible facial expression using FACS (see section 2.1.4). This is achieved by decomposing an expression into the specific AUs and their temporal segments. As AUs are independent of any (cultural) interpretation, they can in turn be used for any higher order decision making process, including the recognition of basic emotions [Ekman et al., 2002], cognitive states like (dis)agreement and puzzlement [Cunningham et al., 2004], psychological states like pain [de C. Williams, 2002], and social signals like emblems (i.e., culture-specific interactive signals like wink, coded as left or right AU46), regulators (i.e., conversational mediators like the exchange of a look, coded by FACS eye descriptor codes for eye position), and illustrators (i.e., cues accompa-

nying speech like raising of the eyebrows, coded as AU1+AU2) [Ekman and Friesen, 1969].

Hence, AUs are extremely suitable to be used as mid-level parameters in an automatic facial behaviour analysis system. They reduce the dimensionality of the problem (a few thousand anatomically possible facial expressions [Cohn and Ekman, 2005] can be represented as combinations of about 60 AUs and Action Descriptors). This representation in AUs can in turn be mapped to any higher order facial expression representation.

It is not surprising, therefore, that automatic AU coding in face images and face image sequences attracted the interest of computer vision researchers. Historically, the first attempts to encode AUs in images of faces in an automatic way were reported by Bartlett et al. [Bartlett et al., 1999], Lien et al. [Lien et al., 1998], and Pantic et al. [Pantic et al., 1998]. These three research groups are still the forerunners in this research field. Two works that could perhaps be regarded as the most influential and successful early works on automatic AU recognition were those by Bartlett et al. [Bartlett et al., 2004] and Tian et al. [Tian et al., 2001]. The work of Bartlett et al. is purely appearance-based. It uses Haar features and boosting techniques to locate the face in an image. The detected face region is then passed through a bank of Gabor wavelet filters. The responses to these filters are fed to a GentleBoost algorithm to select the most important features, which are subsequently used to train Support Vector Machines. The work of Tian et al. is discussed in section 2.2.1 below.

The focus of the research efforts in the field was first on automatic recognition of AUs in either static face images or face image sequences displaying facial expressions produced on command. Several prototype systems were reported that can recognise deliberately produced AUs in either (near-)frontal view [Bartlett et al., 1999, Pantic and Rothkrantz, 2004, Tian et al., 2001] or profile view face images [Pantic and Patras, 2006, Pantic and Rothkrantz, 2004]. Most of these systems employ expert rules and eager learning methods such as neural networks and are either feature-based (i.e., use geometric features like facial points or shapes of facial components) or appearance-based (i.e., use texture of the facial skin including wrinkles and furrows).

One of the main critics that these works received from both cognitive and computer scientists is that the methods are not applicable in real-life situations where subtle changes in facial expression rather than exaggerated AU activations (typical for deliberately displayed facial expressions) typify the displayed facial behaviour. Hence, the focus of the research in the field started to shift to automatic AU recognition in spontaneous facial expressions (produced in a reflex-like manner).

Only recently, a few works have been reported on machine analysis of AUs in spontaneous facial expression data [Bartlett et al., 2006, Cohn et al., 2004b, Valstar et al., 2006]. These methods employ probabilistic, statistical, and ensemble learning techniques, which seem to be particularly suitable for automatic AU recognition from face image sequences [Bartlett et al., 2006, Tian et al., 2005] and are either feature- or appearance-based.

The configuration of facial expressions (either in terms of activated AUs constituting the observed expression or otherwise) has been the main focus of the research efforts in the field. However, both the configuration *and* the temporal dynamics of facial expressions (e.g., the timing and the duration of various AUs) are important for interpretation of human facial behaviour. In fact, the body of research in cognitive sciences which argues that the dynamics of facial expressions are crucial for the interpretation of the observed behaviour, is ever growing [Ambadar et al., 2005, Bassili, 1978, Russell and Fernandez-Dols, 1997].

A complete overview of all the related work published in the field of automatic facial expression analysis would comprise a thesis of its own. Therefore we will only provide a discussion of the most important works with respect to the system described in this thesis. We will focus first on the automatic recognition of Action Units using geometric-feature based techniques in section 2.2.1. Next we will discuss works that specifically address the issue of facial action temporal dynamics in section 2.2.2. Because the small amount of work published in the area of facial action temporal dynamics, we will not limit ourselves to works that address AU temporal dynamics analysis only. Works that deal with spontaneous facial expressions and FACS are treated separately in chapter 9, in which we will also present two studies where the system proposed in this thesis is applied to spontaneous facial expression data.

2.2.1 Geometric feature based FACS detectors

Geometric facial features describe the shapes and locations of facial components such as the eyebrows, eyes, nose, mouth and chin. Based on these facial components, which for example could be described by facial points, feature vectors that represent the face geometry are computed. In the simplest case these would be the locations and shape of the facial components, but more elaborate features, e.g. a parameterisation of the path that a facial point follows over time, could be contrived as well.

The vast majority of efforts to automatically detect AUs have used appearance-based features. It

has even been claimed that appearance-based methods outperform those based on geometric features [Bartlett et al., 1999]. Yet the very first studies on automatic facial action recognition proposed geometric feature-based approaches [Parke, 1974, Platt and Badler, 1981, Suwa et al., 1978]. Also, in this thesis we will show that appearance-based approaches do not always outperform feature-based approaches (see chapter 6).

The first geometric feature based approach was proposed in 1978 by Suwa et al. [Suwa et al., 1978]. They proposed a system that attempts to automatically recognise facial expressions by analysing the positions of 20 characteristic facial points. The points were marked on the face with markers and after filming the coordinates were manually retrieved from the recorded frames. It is noteworthy that they already acknowledged the fact that they had to deal with 3D rigid head motion in order to capture naturalistic expressions and they already proposed a method to separate rigid from non-rigid facial point motion. They modeled expressions in terms of a small number of mid-level parameters which had a close relation to muscle activations, yet they stopped short of expressing them in terms of AUs. However, as FACS has only been introduced in 1976 to an audience of cognitive scientists, it is not surprising that Suwa et al. did not use FACS for their research. On the contrary, it is reassuring to see that they suggested to model muscle actions, independently of Ekman's findings. In their experiments they recognised 5 different emotional activities which were shown by experienced actors on command. They used a modified k-Nearest Neighbour approach, using weights relative to the frequency of the classes.

Terzopoulos and Waters [Terzopoulos and Waters, 1990] presented a system that both synthesised and analysed a face. They used deformable contours (snakes) to track the rigid motion of the head and the non-rigid motions of the eyebrows, nasal furrows, mouth, and jaw in the image plane. The method employed a physically-based synthetic tissue model that models the effects of muscle actions as operations on a lattice of point masses connected by nonlinear elastic springs. As the goal of this paper was only to drive the graphical face model, no emotions or facial muscles were explicitly detected, although the system could presumably be extended to accommodate this. The paper describes a test on a single person performing a single expression while wearing heavy makeup on all the facial features to be tracked. The fact that this system requires that the facial features be highlighted with makeup for successful tracking is a severe limitation. Also, the deformable contours (i.e. snakes) need to be initialised manually in the first frame. The authors note that the 3D geometry of the face model may differ significantly from that of the subject, which could cause difficulty tracking faces whose faces are

very dissimilar to the face model.

Essa and Pentland described in 1997 a system that uses optical flow tracking to fit a face-model mesh to the face [Essa and Pentland, 1997], similar to the approach of Terzopoulos and Waters. The face-model incorporates the modelling of anatomically-based muscles and the elastic nature of facial skin, which have time dependent states and state evolution properties. The 3D mesh is automatically fitted to the face in the first frame of an image sequence and the shape model is updated in each frame using optical flow. The authors proposed a new facial action model based on FACS, called FACS+. They fail to accurately describe what this extended model entails, and exactly how it differs from FACS. The authors never continued their research into this enhanced FACS either. Apart from a couple of pretty pictures and a qualitative discussion of the system, no performance measures were provided.

This is the first paper that utilised optical flow (see [Lucas and Kanade, 1981] for details on the Lucas-Kanade optical flow retrieval method) in a system that aims to detect AUs. Optical flow is a concept which approximates the motion of objects within a visual representation. The difference is that true physical motion in the world of objects is not always reflected in gray value or colour changes in the corresponding images, and gray value changes are not always due to motion of objects. When used as the driving force for object tracking, such as Essa and Pentland proposed, this will lead to a type of error commonly called drift. This type of error occurs for any inference where the state at time t is a result of the optical flow analysed over a long period of time. We therefore find that optical flow is not a useful tool for long-term analysis and should only be used for (semi) instantaneous reasoning.

Lien et al. compared in 1998 a geometry based approach with two appearance based approaches [Lien et al., 1998]. To detect three AUs, the feature based approach tracked 8 facial feature points using optical flow. The first appearance based techniques used a PCA decomposition of the optical flow field on the forehead. The second appearance based technique extracted oriented edges from the forehead using a kernel-based edge detector. On 75 videos taken from the Cohn-Kanade database, the geometric feature based technique achieved an 85% classification rate, the optical flow field approach a 93% classification rate and the oriented edge detection technique an 85% classification rate. All techniques used Hidden Markov Models as classifiers. This system suffers from a number of drawbacks. It needs to be initialised manually, that is, the facial points needed to be manually localised by the operator. It has no 3D model of the face, instead assuming that the face is a plane and that affine transformations would suffice to register the face. A separate classifier was trained for each signal and for each AU combination. As the number of AU combinations lies in the thousands, this approach is

not feasible. Also, the three information signals were not fused in any way, which could have led to significant improvements.

The same lab presented in 2001 a highly successful improvement of this system [Tian et al., 2001]. The manual initialisation remains largely the same, although they introduced a face detector which made it possible to give rough initial estimates of the facial point positions. The proposed method detects AUs separately for the lower face (6 AUs detected) and upper face (10 AUs detected), using Artificial Neural Networks as classifiers. Lip contours, brows, eyelids and 6 points on top of the cheeks are tracked using a variety of highly customised trackers, employing colour histograms, optical flow motion information and shape constraints. Different facial features are tracked with differently operating trackers. The features that function as input to the Neural Networks are expert-defined mid-level parameters computed from the tracked facial points, such as the distance between the eyelids, plus edge information taken from the regions next to the eyes, on the bridge of the nose and next to the nose (the nasolabial furrows). Surprisingly, the authors chose not to use the optical flow field anymore, even though they reported that this feature was the most successful in their previous study. The work reported an unprecedented thorough performance evaluation. The system was tested on over 230 image sequences taken from the Ekman-Hager database, and separately on over 200 examples taken from the Cohn-Kanade database. As a third test, to evaluate how well the system could generalise to completely novel data, they trained their system on one database and tested it on the other. For this test they achieved an incredibly high 93% classification rate.

Gokturk and colleagues proposed a method that simultaneously tracks the 3D head pose and face shape [Gokturk et al., 2002]. A 3D shape model consisting of 19 points is initialised manually in the first frame of an image sequence and tracked using optical flow. The system utilises Support Vector Machines (SVMs) to recognise 5 distinct facial actions: the neutral expression, opening/closing of the mouth, a smile and raising the eyebrows. They reported a mean 91% classification accuracy, however, the system was tested on data from only three persons. The data of two of these subjects were used to both train the classifier as well as test it, making the results extremely subject dependent. Overall, the results are inadequate to judge the true performance of this system.

Zhang and Ji [Zhang and Ji, 2005] proposed a hybrid approach that uses both appearance and geometric features. They use an active infrared illuminator, making use of the red-eye effect to detect the pupils. According to the paper, they use the locations of the pupils together with anthropometric statistics to localise the regions of interest in which the facial features can be found. Next they rep-

resent the image by passing it through a filterbank of 18 Gabor filters. Unfortunately that is all they say about the way the facial points are detected, leaving an important part of their work undiscussed. These facial points are tracked in the wavelet domain using Kalman filtering. As appearance features they use the classical Canny edge detection method, applied in a number of regions where furrows are to be expected⁶. Why the authors take the extra step of using the Canny method for oriented edge detection, instead of re-using the Gabor filters, which can be used as oriented edge detectors, remains unexplained.

The authors propose expert rules for the relations between AUs and the geometric properties of the tracked points, claiming that expert rules are more robust than machine learning methods. This claim is not backed up by qualitative nor by quantitative results. The method features a complex temporal dynamic probabilistic classification strategy based on Dynamic Bayesian Networks. From the geometric and appearance based features, they first model the probability that an AU is activated. In a layer between AUs and emotions, they add nodes that group AUs in two: the first group of AUs are considered ‘typical’ for that emotion, while the second group of ‘auxiliary’ AUs are less important, according to an expert’s opinion. While the authors claim to address many open issues of facial expression analysis with this complex system, many assumptions and system design choices are ill-chosen in my opinion. It comes as no surprise to me that the authors provide no qualitative performance results of their AU and emotion detection system, other than the emotion detection on a single image sequence containing each emotion exactly once, one after the other, separated by neutral expressions.

Kotsia and Pitas recently proposed to use a system very similar to the one presented in this thesis [Kotsia and Pitas, 2007]. They manually fitted a candid face mesh consisting of 79 nodes to the first frame. The authors noted that usually manually specifying the positions of 5-8 nodes was enough for a good fit of the facial mesh. This mesh is subsequently tracked through the image sequence using Kanade-Lucas-Tomasi tracking [Lucas and Kanade, 1981], and node-displacement features are computed. SVMs were used to recognise emotions and 17 AUs. On the Cohn-Kanade database they achieved a classification rate of 93.5%.

⁶Furrows in the face due to facial expression can be expected outside the eyes (crow-feet wrinkles), on the bridge of the nose, between and above the eyebrows, left and right of the nose, on the chin-boss, and left and right of the mouth corners

2.2.2 Temporal dynamics analysis

Facial expression temporal dynamics are essential for categorisation of complex psychological states like various types of pain and mood [de C. Williams, 2002]. They are also the key parameter in differentiation between posed and spontaneous facial expressions [Ekman and Rosenberg, 2005]. For instance, it has been shown that spontaneous smiles, in contrast to posed smiles (e.g. a polite smile), are slow in onset, can have multiple AU12 apices (multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 withing 1s (see section 9.3, [Cohn and Schmidt, 2004]). Similarly, it has been shown that the difference between spontaneous and deliberately displayed brow actions (AU1, AU2, AU4) lies in the duration and the speed of onset and offset of the actions and in the order and the timing of the occurrences of facial actions (see section 9.2).

In spite of these findings, the vast majority of the past work in the field does not take dynamics of facial expressions into account when analysing the shown facial behaviour. Some of the past work has used aspects of temporal dynamics of facial expression such as the speed of a facial point displacement or the persistence of facial parameters over time. However, this was either done to increase the performance of the accuracy with which static aspects of facial expressions were detected (i.e., AU activation) [Gralewski et al., 2006, Tong et al., 2006, Zhang and Ji, 2005] or in order to report on the intensity of (a component of) the shown facial expression [Littlewort et al., 2006, Zhang and Ji, 2005]. Hardly ever were the properties of facial expressions' temporal dynamics explicitly analysed.

Exceptions from this overall state of the art in the field include two studies on automatic segmentation of AU activation into temporal segments (neutral, onset, apex, offset) that were made in parallel with the work presented in this thesis, in the same lab. Temporal segments were recognised in frontal- [Pantic and Patras, 2005] and profile-view [Pantic and Patras, 2006] face videos. Both of these works employ geometric features and rule-based reasoning to encode AUs and their temporal segments.

Bartlett et al. [Bartlett et al., 2003] report that, although their SVM classifiers were not trained to measure the amount of eye opening, they learned this as an emergent property. What they reported was that the output of a SVM is an indication of the strength of a facial action. This, however, is not completely true. An SVM learns a decision boundary based on data elements close to data elements of the opposite class. Essentially, it learns the decision function based on the atypical examples, rather than the typical examples. Because of this, it is not correct to assume that having a greater distance

to the decision boundary means that a test example is more typical for that class, or, as the authors put it, has a higher intensity. That being said, the output of the classifiers does appear to be a smooth function over time. However, the output has many peaks (more than there should be in the expression shown) and the graphs shown do not seem to correspond with actual temporal segments of facial actions. The authors do not report on any qualitative or quantitative fitness of the SVM output with the facial action intensity.

Apart from reporting on the SVM outputs, the authors use Hidden Markov Models (HMMs) to increase the robustness of their AU activation detection system. The HMMs learn the temporal persistence of facial actions and therefore increase classification accuracy. The system was able to detect three AU combinations: AU1+AU2 (brow raiser), AU4 (brow lowerer) and AU45 (a blink). The brow action classifier was implemented as a forced-choice classifier with three classes (neutral, AU4 and AU1+AU2) and achieved a reported 75% classification rate. The blink detector was implemented as a binary detection problem with a reported performance of 98.2% classification rate.

A structured approach to dealing with the particularities of the temporal dynamics of facial expressions was proposed by El Kaliouby and Robinson [Kaliouby and Robinson, 2004]. The authors point out that there are different levels of temporal abstraction to facial expressions. They model three levels of abstraction: at the highest level there are mental states which typically last between 6-8 seconds, then there are facial displays which last up to 2 seconds, and at the lowest level there are the action units that last tenths of seconds. This last statement seems a little arbitrary to me. They do not mention what they base this duration on and many AU activations present in publicly available facial expression databases last several seconds (see chapter 3 for an overview of facial expression databases). Still, the approach seems logical.

The proposed system employs Dynamic Bayesian Networks, of which HMMs are a subclass, to model both the evolution of time and the relationships between the three abstraction levels. Overall, the system seems a good approach to complex facial action understanding. Yet their system is still only an activation detection system, and it does not gather any temporal dynamics information which could be used by subsequent facial analysis systems. Also, the systems used to initialise the locations of facial points, track the points in subsequent frames and compute features for facial expression analysis are not entirely state-of-the-art. They employ techniques such as colour histograms, motion flow and a number of head pose estimation techniques that seem incredibly sensitive to noise.

The system proposed by Zhang and Ji [Zhang and Ji, 2005] was already mentioned in section 2.2.1 above. Thus let us describe the temporal dynamic aspects of that system here. A Bayesian network is learned to recognise an emotion at each moment in time. At the lowest level there is a layer of visual cues such as whether the mouth opening has exceeded a certain threshold or not. These are connected to a layer of nodes representing active AUs. These are in turn connected to a layer of emotion classes. At this highest level an HMM connects the emotion predictions through time. This allows for smooth transitions from one emotion to another, e.g. from surprise to happiness, or from an emotive expression to neutral. This system achieves two (related) objectives: it increases the activation detection performance of emotions, and it models which emotions are likely to follow after each other. A follow-up paper was presented by the same group in which they use the same approach, but now with the aim to recognise the AU activations instead of emotions [Tong et al., 2007]. Thus, in the top level, not emotions but AU activations are connected by the HMMs.

The title of the paper by Littlewort et al. [Littlewort et al., 2006], ‘Dynamics of facial expression extracted automatically from video’ sounds very promising, but it is actually the least interesting of the papers described here, with regard to facial dynamics. For recognition of the 6 Ekmanian emotions they employ the same Gabor filter features as presented in their previous work [Bartlett et al., 2003], which themselves can not encode any temporal dynamics information. They experiment with a number of classifiers, none of which can model time. The HMM, employed in the authors’ previous work, was not one of the evaluated classifiers, while that classifier *is* capable of modelling temporal dynamics. Basically, the *dynamics* part of their title is justified by their observation that the output of their SVMs is a smooth function over time. This is the same claim as was made earlier in [Bartlett et al., 2003]. Again, this statement is made after human visual inspection of the SVM output, no quantitative evaluation of this claim is provided.

The first effort to explicitly analyse the characteristics of facial temporal dynamics themselves was presented by Pantic and Patras in 2005 [Pantic and Patras, 2005], with a follow-up paper that employed the same techniques yet applied to profile-view images in 2006 [Pantic and Patras, 2006]. In these works, the authors propose to segment a facial action in temporal phases. A facial action can be divided into four temporal phases or segments: the onset, apex, offset and neutral phase of a facial action. For an in-depth discussion of this principle the reader is referred to section 7. The system proposed by Pantic and Patras was developed in parallel with the system described in this thesis. It employs the same particle-based facial point tracker as described in section 5.1. Based on the tracked

points data, three feature functions are computed; the deviation of a point's x -position relative to a neutral position, its y -value deviation relative to a neutral position, and for combinations of two points at a time whether the distance between the points has increased, decreased or remained the same. The functions have a discrete output: for the first function the output is either {left, no motion, right}, for the second {up, no motion, down} and for the third {increase, no motion, decrease}. Both functions use a threshold that interestingly has the same value for all functions, for all AUs to be recognised. The same threshold value is used both for AU activation detection (i.e. the detection of the presence of an AU, regardless of its temporal phase) as for temporal segment detection.

An interesting observation is that all systems mentioned above rely purely on the temporal modelling powers of the classifiers. None particularly attempt to capture information about the temporal dynamics in the description of their features. The most powerful temporal dynamics descriptors used are the difference of a parameter value compared to its value at an earlier, often neutral state of the face. Relying on the classifiers only seems a naive approach, as they can only perform so well given the data they are presented with. Designing complex multi-stage classifiers that capture temporal dynamics is an active and promising research area. However, it seems strange to focus solely on the classifier part to model the temporal dynamics while it is relatively easy to encode information on the temporal dynamics of facial actions directly in the features.

Chapter 3

Facial expression databases

To be able to evaluate the geometric-feature based AU analysis system that we will propose in this thesis, we need a suitable set of data. We have chosen two databases to use in our studies. In this chapter we will outline our reasons for this choice. We will first give an overview of the existing facial expression databases. We will then give a detailed description of the two databases that are most appropriate to evaluate our methods. The first database we chose is the DFAT-504 Cohn-Kanade database [Kanade et al., 2000]. We will show that the DFAT-504 Cohn-Kanade database lacks a number of properties needed to evaluate all the aspects of our system and therefore we introduce in this chapter the MMI-Facial Expression Database [Pantic et al., 2005b], which was purposefully created to evaluate our proposed facial expression analysis system.

A second reason to create the MMI-Facial Expression Database is to provide the facial expression recognition community with a publicly available database, so that novel facial expression recognition techniques can be tested on a large, common set of data. It has been noted in the literature that to develop, evaluate and compare automatic facial expression recognition systems, large collections of training and test data are needed [Kanade et al., 2000, Pantic and Rothkrantz, 2003]. The establishment of an easily accessible, comprehensive, database of facial expression exemplars has become an acute problem that needs to be resolved if fruitful avenues for new research in automatic facial expression analysis are to be opened.

In this chapter we will discuss the desired characteristics of a facial expression database that is to be used as a common resource for scientists in the field. We will then give an overview of existing facial expression databases. We will give an in-depth evaluation of the first database we use to evaluate our

systems on, the DFAT-504 Cohn-Kanade database. This is the most used database for automatic facial expression recognition evaluation to date. We will explain why it lacks some of the currently desired properties of a benchmark database. Finally, we will present the MMI-Facial Expression database, which was developed to facilitate the current needs of facial expression analysis researchers.

3.1 Facial expression database characteristics

Below we will discuss what are the important characteristics of a facial expression database. The quality of a comprehensive facial expression database depends primarily on the data contained therein. We will then provide an overview of the existing facial expression databases.

3.1.1 Desirable data characteristics

Static images and videos

While videos are necessary for studying temporal dynamics of facial behaviour, static images can be used to obtain information on the configuration of facial expressions [Pantic and Rothkrantz, 2003] which is essential, in turn, for inferring the related meaning of an expression (e.g. in terms of emotions). Of course one could argue that it is possible to extract still images from videos. However, because of the limitations of current capture and storage devices, videos are often recorded at relatively low spatial resolution. This puts limitations on systems that process videos. As static images only record one moment in time, their storage and capture requirements are less constricting. This in turn allows a facial expression database to include a large number of these high-resolution still images, which can be used to perform studies that require a higher detail of the facial features. Therefore a good benchmark database should include both high resolution static images and videos of faces showing prototypic expressions of emotion and various expressions of single AU and multiple AU activation.

Data descriptors

A fundamental factor in designing a benchmark database is the metadata that are to be associated with each database object. These metadata provide a description of the characteristics of the data objects in a human readable form, which makes it possible to discriminate between groups of data.

These data could be used as the ground truth in the evaluation of the performance of automated facial expression analysers or for the selection of a subset of data with specific properties.

For general relevance, the imagery should be scored in terms of AUs and their temporal segments since FACS provides an objective and comprehensive language for describing facial expressions and a general representation that can be used in many applications (see section 2.1). However, as many researchers still propose systems to distinguish the six basic emotions, and the interpretation of the meaning of the expression is arguably ultimately the goal of any facial expression recognition system, such information should be available as well. However, as this interpretation label is highly subjective, it might very well be that an expression is described using multiple, possibly many, different labels.

In general, there should be a facial expression interpretation label (e.g. basic emotion or another description of an expression such as ‘boredom’ or ‘sleepy’) as well as FACS coding. Note that it takes more than one hour to manually score 100 still images or a minute of videotape in terms of AUs and their temporal segments [Ekman et al., 2002]. Hence, obtaining large collections of AU-coded facial-expression data is extremely tedious and expensive, and, in turn, difficult to achieve.

Other metadata to include are information about the gender, age and ethnicity of the subjects, whether there are any occlusions present and information about the context in which these recordings were made such as the utilised stimuli, the environment in which the recordings were made, the presence of other people, etc. [Pantic and Bartlett, 2007]. For increased flexibility, evolution of the data description and ease of use, metadata is provided to the user in XML format.

Individual differences between subjects

A very important variable of the benchmark database is the variety of the subjects included. Features such as hair style, hair colour, glasses or skin colour could strongly affect the performance of the facial expression analysis. Face shape, texture, colour, and facial and scalp hair vary with sex, ethnic background, and age [Farkas and Munro, 1987]. Beards, glasses, make-up or jewellery may obscure, occlude or enhance facial features. In order to develop algorithms that are robust to individual differences in facial features and behaviour, it is essential to include a large sample of subjects who vary relative to the values of the above-mentioned parameters.

Variable head pose

One of the major obstacles encountered at this moment by researchers in the field is the variability in the head pose in an initial image as well as over time. To study methods to tackle this problem, a comprehensive facial expression database should contain a large number of videos in which the head pose varies over time, and/or which are recorded by multiple cameras from different angles. This way, relations between camera viewing angles, head pose, facial features and facial expression can be learned. The extremes of the camera viewing angle are the frontal view (camera right in front of a person), the profile view (camera to the left or right of a person), a view from a very low camera (possibly meters lower than the observed face) and the helicopter view (camera directly above a person). The use for the latter two views is primarily for applications such as surveillance, where a (mobile) camera might be positioned at a different height than the observed face is.

There are other reasons to use multiple cameras when recording the face. For instance it would be easier to reconstruct a 3-D model of the face with such data [Park and Jain, 2006]. Also, symmetry properties of a facial expression could be studied in greater detail, which could reveal the intent with which the expression was made [Ekman and Friesen, 1974], (see chapter 9).

Posed and spontaneous actions

Facial expressions are either elicited on purpose as a voluntary action or unconsciously, that is, spontaneously. These two types of expression vary widely in terms of configuration and temporal dynamics (see chapter 9 for a thorough discussion on this topic). A very important issue for the creation of a database is that it should consist of both deliberate (posed), actions performed on request as well as spontaneous actions that are not under volitional control of the subject.

Volitional facial movements originate in the motor cortex, whereas the involuntary facial actions, originate in the sub-cortical areas of the brain [Ekman and Friesen, 1974]). Therefore, some facial movements might be easier to make deliberately compared to others. In particular, while few people are able to perform certain facial actions voluntarily (e.g., AU2), many are able to perform these actions spontaneously [Kanade et al., 2000]. This may make it hard to collect posed expression data of all possible AUs. On the other hand some expressions such as fear or anger might prove very difficult to illicit spontaneously from a subject without breaking ethical codes of conduct.

Configurational and dynamic aspects

The dynamics and the configuration of a facial expression complement each other and are both necessary in analysing a facial expression. Virtually all the existing facial expression analysers assume that the input expressions are isolated or pre-segmented, showing a single temporal activation pattern (neutral \implies onset \implies apex \implies offset) of either a single AU or an AU combination that begins and ends with a neutral expression (see chapter 7 of this thesis). In reality, facial expressions do not necessarily evolve with this particular sequence of temporal phases. Facial expressions are very complex and the transition from an action or combination of actions to another might not follow this pattern and might not involve an intermediate neutral state. Among other things, these temporal dynamics are a key parameter in differentiating posed from spontaneous facial displays (see chapter 9).

Examples of both the neutral-expressive-neutral and the variable-expressive-variable behaviour should be included in a good database in order to study the essential question of how to achieve parsing of the stream of behaviour. Facial expression temporal dynamics form a critical factor in applications that have to do with social behaviour, such as social inhibition, embarrassment, amusement and shame [Pantic and Bartlett, 2007].

Facial expression intensity

Facial expressions vary in intensity. As defined by Pantic & Bartlett [Pantic and Bartlett, 2007]:

By intensity we mean the relative degree of change in facial expression as compared to a relaxed, neutral facial expression. In the case of a smile, for example, the intensity of the expression can be characterised as the degree of upward and outward movement of the mouth corners, that is, as the degree of perceivable activity in the Zygomaticus Major muscle (AU12) away from its resting, relaxed state.

It has been experimentally shown that the expression decoding accuracy and the perceived intensity of the underlying affective state vary linearly with the physical intensity of the facial display [Hess et al., 1997]. Hence, explicit analysis of expression intensity variation is very important for accurate expression interpretation. It is also an important factor in distinguishing between spontaneous and posed facial behaviour (see chapter 9). Hence, a comprehensive facial expression database should contain examples with a wide range of intensities. How the intensities should be measured is an

entirely different matter, but this need not be addressed when creating the database. An intensity measure label could be added at any stage.

3.1.2 Data collection issues

Several technical considerations for the database should be resolved including field of sensing, spatial resolution, frame rate, data formats, and compression methods. The choices should enable sharing the data between different research communities all over the world, on any existing and (where possible) future operating system. Database creators should try not to make any assumptions about how the data will be used in the future by scientists; they should not throw away any information, even if they deem that information not important at the time of recording. Any compression performed should therefore be lossless. This severely limits the possibilities for data formats and compression techniques.

For the acquisition of posed actions, performed on request, the subjects should be either experts in production of expressions (e.g., trained FACS coders or actors) or individuals being instructed by such experts on how to perform the required facial expressions. Given the large number of expressions that should be included into the database, provisions should be made for individual researchers to append their own research material to the database. However, a secure handling of such additions has to be facilitated.

3.1.3 Database accessibility and usability

Benchmark databases of facial expression exemplars could be valuable to hundreds of researchers in various scientific areas if they would be easy to access and easy to use. It would be ideal to have a relaxed level of security, which allows any user a quick, web-based access to the database and frees administrators of time-consuming identity checks.

However, in this case, non-scientists such as journalists and hackers would be able to access the database. If the database is likely to contain images that can be made available only to certain authorised users (e.g., images of psychiatric patients), then a more comprehensive security strategy should be used. For example, a Mandatory Multilevel Access Control model could be used in which the users can get rights to use database objects at various security levels (e.g., confidential, for internal use only, no security).

Name	No. Subjects	No. Videos	No. Images	Available	Web-based	Searchable
Adult Attachment Interview [Roisman et al., 2004]	64	64	0	No	No	No
AMI [McCowan et al., 2005]	12	265	0	Yes	Yes	No
AR [Martinez and Benavente, 1998]	126	0	4000	Yes	Yes	No
AT&T (formerly ORL) [Samaria and Harter, 1994]	30	0	400	No	No	No
Belfast Db [Douglas-Cowie et al., 2003]	125	239	0	Yes	No	No
BU-3DFE [Yin et al., 2006]	100		2500 3D models	Yes	No	No
CVL [Solina et al., 2003]	10	0	798	No	No	No
Cohn-Kanade [Kanade et al., 2000]	97	488	0	Yes	No	No
FG-NET [Wallhoff, 2006]	18	399	0	Yes	Yes	No
JAFFE [J. et al., 1998]	10	0	219	Yes	Yes	No
MMI [Pantic et al., 2005b]	90	2363	740	Yes	Yes	Yes
PIE [Sim et al., 2003]	68	0	40000	Yes	No	No
RU-FACS [Bartlett et al., 2006]	100	100	0	No	No	No
SAL	2	10 hrs	0	Yes	No	No
UT Dallas [O’Toole et al., 2005]	284	800	0	Yes	No	Yes
Yale Face Db-part B [Georghiades et al., 2001]	10	0	5850	Yes	No	No

Table 3.1: Existing facial expression databases

3.2 Overview of existing databases

There have been a number of databases created over the past few decades, none of which addresses all of the above issues successfully. Tables 3.1 and 3.2 list a number of important efforts. In the tables we have included information on 16 facial expression databases that have been used in the literature on facial expression recognition. Only databases that are freely available to the research community have been included. We have listed a number of characteristics that we deem important: the number of subjects, videos and images contained in the database as well as whether the database is publicly available (3.1). The requirement to be listed as being web-available is that the material is downloadable and that there is a publicly available link to start the download (see column 6 in table 3.1). So, if the website of a database would only provide some information about the database and state that interested parties should contact one of the authors, this would not count as being web-available.

The column on searchability lists whether means to search through the database’s content are readily available. In other words, tools should be available that allow the user to fill in a number of criteria and run the search for the user. This does not necessarily have to be integrated in an online web

Name	Expression description	P/S	Occlusions	Views
Adult Attachment Interview	FACS, 6 basic emotions, negative/positive affect	S	None	Frontal
AMI	1 basic emotion	S	Glasses	Frontal to profile
AR	3 basic emotions	P	Glasses, scarf	Frontal
AT&T	2 basic emotions	P	Glasses	Frontal
Belfast Db	Multiple emotive	S	Unknown	Frontal
BU-3DFE	6 basic emotions	P	None	Frontal, 45, -45
CVL	smile/neutral	P	None	Frontal, 2 profiles, 45, -45
Cohn-Kanade	6 basis emotions, FACS	P	None	Frontal
FG-NET	6 basic emotions	S	None	Frontal
JAFFE	6 basic emotions	P	None	Frontal
MMI	6 basic emotions, FACS	P,S	Glasses, facial hair	Frontal, profile
PIE	smile/neutral/blink/talk	P	None	13 views
RU-FACS	FACS	S	Unknown	Frontal, 2 profiles
SAL	FEEL-TRACE	S	Unknown	Frontal
UT Dallas	6 basic emotions + puzzlement/laughter/boredom/disbelief	P	None	Frontal
Yale Face Db-B	1 basic emotion	P	None	9 views

Table 3.2: Characteristics of facial expression databases

interface, but it should be integrated with the database.

Table 3.2 lists how the expressions are described, i.e. what the available metadata are such as whether the data was posed (P) or spontaneous (S), what types of occlusions were present and from what views the imagery was recorded.

The AMI database is actually much larger than listed in table 3.1. However, most of those data are not annotated for facial expression at all. There is only annotation for the presence of laughters/smiles of 265 videos. For that set of videos we have listed the information in table 3.1.

In the following sections, we will only discuss the most popular database: the Cohn-Kanade database and the MMI-Facial Expression Database, the creation of which was a part of the work done for this thesis. We do so because it is these two databases that we use in this thesis to evaluate our techniques. The reader is referred to [Cowie et al., 2005, Gross, 2005, Zeng et al., 2008], in which most of the databases listed above are analysed in great detail.

3.3 Cohn-Kanade database

The Cohn-Kanade DFAT-504 facial expression database was developed for research in recognition of the six basic emotions and their corresponding AUs. The database contains 488 near frontal-view videos of facial displays produced by 96 adults aged 18 to 50 years old. Of the subjects, 69% is female,

81% is Caucasian, 13% African and 6% of the subjects are from other ethnic groups. The database was made publicly available as a set of grey-scale image sequences, that start with a neutral expression and end as soon as the peak expression (apex) is reached (i.e. not the entire apex phase is included). The publicly available DFAT-504 database is a subset of about 2000 originally recorded videos. The original videos, in colour and displaying the full neutral-expressive-neutral expression are still kept at the authors' lab, but are not publicly available.

This database is currently the most commonly used database for studies on automatic facial expression analysis. All facial displays were made on command and the recordings were made under constant lighting conditions. Two certified FACS coders provided AU coding for all videos. Inter-observer agreement was expressed in terms of Cohens kappa coefficient, which is the proportion of agreement above what would be expected to occur by chance. The mean kappa for inter-observer reliability was 0.82 for AUs at apex.

There are a number of reasons why the DFAT-504 database is not suitable for training or testing algorithms that intend to address the current challenges of automatic facial expression recognition. First of all, the variability of facial actions within the database is very low. Not only because the database was recorded with the intent of recognising only the six basic emotions, but also because most subjects depict the same emotion in the same way as the other subjects did. Of course the number of ways in which the six basic emotions can be displayed is limited, yet the number of variations shown in this database is even lower. This is probably because the certified FACS coders showed the subjects how to do a typical expression of emotion, and the subjects mimicked that expression. The database therefore does not contain all existing AUs. Also, there exists a high correlation between groups of AUs, which would allow a system to recognise AUs by inference rather than based on observations that are characteristic for the relevant AU only. It is therefore not a suitable source of data to train a system to detect AUs in any real world applications other than basic emotion detection, as in those situations we can expect all AUs to occur in all possible combinations.

Incomplete apex phases and unavailable apex-offset and offset-neutral transitions in the recordings make this database unsuitable for training and testing systems capable of analysing the temporal behaviour of facial expressions. This, in turn, renders the database useless if the goal of a system is a more detailed analysis of a human's facial actions. The database consists exclusively of posed expressions, which do not represent all real-life facial actions. A separate database is needed to learn the characteristics of spontaneous facial expressions.

The videos in the database were recorded from a frontal-view of the face only and contains a negligible amount of head motion. This does not reflect natural human behaviour and therefore the database can not be used to train or test a system that is to function under conditions where the head pose is unconstrained. As a last and final drawback, many recordings have a time-stamp overlayed at the position of the chin, or on the position of the neck just below the chin, which makes it virtually impossible to detect any actions that involve the chin boss or the jaw (e.g. the dropping of the jaw or the sideways movement of the jaw).

3.4 MMI Facial expression database

The MMI-Facial Expression Database, from now on MMI database, has been developed to address most (if not all) of the issues mentioned above. It contains more than 3000 samples of both static images and image sequences of faces in frontal and in profile view displaying various facial expressions of emotion, single AU activation, and multiple AU activation. It has been developed as a web-based direct-manipulation application, allowing easy access and easy search of the available images [Pantic et al., 2005b].

The properties of the database can be summarised in the following way:

Sensing: The static facial-expression images are all true colour (24-bit) images which, when digitised, measure 720*576 pixels. There are approximately 600 frontal and 140 dual-view static facial-expression images (see Fig 3.1). Dual-view images combine frontal and profile view of the face, recorded using a mirror. All video sequences have been recorded at a rate of 25 frames per second using a standard PAL camera. There are approximately 335 frontal-view, 30 profile-view, and 2000 dual-view facial-expression video sequences (see Fig. 3.2). The sequences are of variable length, lasting between 40 and 520 frames, picturing neutral-expressive-neutral facial behaviour patterns.

Subjects: Our database includes 69 different faces of students and research staff members of both sexes (44% female), ranging in age from 19 to 62, having either a European, African, Asian, Carribean or South American ethnic background.

Samples: The database consists of four major parts: one part of posed expression still images, one part of posed expression videos and two parts of spontaneous expression videos. For the posed expressions the subjects were asked to display 79 series of expressions that included either a single AU (e.g., AU2)



Figure 3.1: Examples of static frontal-view images of facial expressions in the MMI Facial Expression Database



Figure 3.2: Examples of apex frames of dual-view image sequences in MMI Facial Expression Database

or a combination of a minimal number of AUs (e.g., AU8 cannot be displayed without AU25) or a prototypic combination of AUs (such as in expressions of emotion). The subjects were instructed by an expert (a FACS coder) on how to display the required facial expressions.

For the posed videos, the subjects were asked to include a short neutral state at the beginning and at the end of each expression. They were asked to display the required expressions while minimising out-of-plane head motions. The posed expression videos were recorded using a mirror placed on a table next to the subject at a 45 degrees angle so that the subject's profile face was recorded together with the frontal view. The posed expression videos were recorded with a blue screen background and two high-intensity lamps with reflective umbrellas (e.g., Fig. 3.2).

The first part of the spontaneous expression videos was recorded in a living room environment with no professional lighting added. The subjects were shown both funny and disgusting clips on a PC. Their reaction to the clips was recorded and cut into separate neutral-expressive-neutral videos. The expressions recorded were mainly happiness and disgust, with a fair number of surprise expressions captured as well.

The second part of the spontaneous data consists of 11 primary school children who were recorded by a television crew of the Dutch tv program ‘Klokhuis’. They were asked to laugh on command. There was also a comedian present who made jokes during the recording, which made the children laugh. These spontaneous expressions were again cut into videos that contain neutral-expressive-neutral sequences. This set of data contains severe head pose variations, as the children were free to move and thought they weren’t being recorded at the time. We will refer to the spontaneous part of the database as the MMI-database part 2 in the remainder of this thesis.

Metadata: Two experts (FACS coders) were asked to code the AUs displayed in the images constituting the MMI Face Database. In the case of a number of facial-expression video sequences, the coders were also asked to depict the temporal segments of displayed AUs. To date, all the objects in the database have been event-coded for overall activation of AUs and approximately 300 videos have their temporal segments of target actions coded frame by frame.

Since human observers may sometimes disagree in their judgements and may make mistakes occasionally, the inter-observer agreement is again measured. Inter-observer agreement was expressed in terms of Cohen’s kappa coefficient [Cohen, 1960], which is the proportion of agreement above what would be expected to occur by chance. The mean kappa for inter-observer reliability was 0.91 for AUs at apex. For the meta data available to the public, when a disagreement was observed the FACS coders made a final decision on the relevant metadata by consensus. As the expressions shown were made on command, there was no need for the subject to self-report on the full expressions (e.g. ‘happiness’, ‘bored’, etc.) that we recorded of them. For the same, measuring the correlation between the requested full expression and the coders decision is not particularly useful and has not been attempted.

Retrieval and Inclusion of Samples: In order to allow a quick and easy web-based access to the database and yet shield parts of the database that can be made available only to certain authorised users, we used a Mandatory Multilevel Access Control model. In this model, the users can get rights to use database objects at various security levels (i.e., administrators, confidential, users). This way it is possible to create different sets of data that are available to different groups of users, such as a set of data for internal use only and a set available to all registered users.

The database allows easy search according to a number of criteria such as data format, facial view, AUs displayed, emotions shown, gender, age, etc. The provision of a preview per sample makes the search even easier. The user can inspect previews of the selected data before deciding to download

the desired material.

The current implementation of the database allows easy addition of either new samples or entire databases. However, only the database administrators can perform such additions; an automatic control of whether an addition matches the specified formats defined for the database objects has not been realised yet.

The database has been developed as a user-friendly, direct-manipulation application. The Graphical User Interface of the web-application is easy to understand and to use. On-demand on-line help and a flash animation demonstrating the path that the user should follow to download the required facial expression exemplars are other elements of user friendliness.

To summarise, the MMI-Facial Expression Database contains examples of all possible AUs and a great amount of examples of the six basic emotions plus a number of other expression categories such as boredom or an expression of ‘I don’t know’. It contains both posed and spontaneous expressions of a large number of people, with occlusions such as beards, headscarfs, glasses and moustaches apparent in great abundance. The spontaneous data contain large in-plane and out-of-plane head motion. All videos are recorded as a full neutral-expressive-neutral expression sequence so that an analysis of the temporal dynamics of facial expressions is possible. All examples are recorded in colour and for a large portion of the database there are synchronised videos of frontal and profile view faces.

Together this makes the MMI-facial expression database the most comprehensive facial expression database available at this moment. The database is already a great success: in 2006 the site was visited 2876 times and in 2007 it was visited 5388 times. 593 users have registered with the database, of which 247 users have proceeded to sign the End User License Agreement, which is a requisite to make search queries and download the data. In the period of March 2006 until December 2007 an average of over 10 GigaByte per month has been downloaded. Many researchers have already used the data for their research and have referred to the database in their publications. For the near future a facial expression recognition competition is planned based on the data from the MMI-facial expression database.

Chapter 4

Facial feature point detection

An automatic facial expression analysis system usually consists of three consecutive steps: data acquisition, feature extraction and classification. In the system we propose here, the data acquisition is performed by one video camera. This means we will be working on time sequences of 2-dimensional images, also called image sequences. From these image sequences we compute geometric features. For the AU and emotion recognition systems, these features will be computed per frame. Later, for analysing complex human behaviour, we will compute temporal-dynamic features that are extracted from a variable number of frames, depending on the analysis of the temporal dynamics of AUs (see chapter 9). The geometric features are then used as input to our machine learning techniques. In this thesis we will propose to use a combination of GentleBoost [Friedman et al., 2000], Support Vector Machines [Vapnik, 1995] and Hidden Markov Models [Rabiner, 1989].

As stated above, our proposed facial expression analysis system will use geometry-based features, extracting such features as the velocities and positions of points and the distances between points. Therefore the first step in our fully automated facial expression recognition system is to find the coordinates of those facial points in each frame of a video. The locations of these points are found in the first frame using a point detector which locates 20 facial points. The 20 points that we detect are shown in figure 4.1 and were chosen as the minimum number of points theoretically needed to distinguish the 26 AUs that can be recognised using geometric features. In all frames after the first we will find the positions of these 20 points using a point particle-filtering based tracker, that uses the coordinates of the points found in the first frame to initialise the tracker. In this chapter we will describe how our facial feature point detection works.

Previous methods for facial feature point detection could be classified into two categories: texture-based and shape-based methods. Texture-based methods model the local texture around a given feature point, for example the pixel values in a small region around a mouth corner. Shape-based methods regard all facial feature points as a shape, which is learned from a set of labelled faces, and try to find the proper shape for any unknown face. Typical texture-based methods include gray-value, eye-configuration and neural-network-based eye-feature detection [Reinders et al., 1996], log-Gabor filter based facial point detection [Holden and Owens, 2002], and two-stage facial point detection using a hierarchy of Gabor filter networks [Feris et al., 2002].

Typical shape-based methods include detectors based on active shape or active appearance models [Hu et al., 2003, Yan et al., 2003]. A number of approaches that combine texture- and shape-based methods have been proposed as well. Wiskott et al. [Wiskott et al., 1997] used Gabor jet detectors and modeled the distribution of facial features with a graph structure. Cristinacce and Cootes used Haar features with an AdaBoost classifier and combined the classifier output with statistical shape models in a probabilistic framework [Cristinacce and Cootes, 2003]. Chen et al. proposed a method that applies a boosting algorithm to determine facial feature point candidates for each pixel in an input image and then uses a shape model as a filter to select the most probable position of feature points [Chen et al., 2004]. They used greyscale image patches as input to the classifiers. Cristinacce and Cootes detected 22 points in frontal faces [Cristinacce and Cootes, 2006]. Like Chen et al. they used greyscale image patches. In a test image the closest K matches for each facial point were found based and these were input to a 2D shape model to find the set of patches that resulted in the most probable positions of the facial points.

In general, although some of these detectors have been reported to perform quite well when localising a small number of facial feature points such as the corners of the eyes and the mouth, none of them detects all 20 facial feature points illustrated in Fig. 4.1 and, more importantly, none performs the detection with high accuracy. To wit, the current approaches usually regard the localisation of a point as a *success* if the distance between the automatically labelled point and the manually labelled point is less than 30% of the true inter-ocular distance (IOD, the distance between the eyes). However, 30% of the true inter-ocular value is at least 30 pixels in the case of the Cohn-Kanade database samples, which we used to test our method. This means that a bias of 30 pixels for an eye corner would be regarded as *success* even though the width of the whole eye is approximately 50 pixels. This is unacceptable in the case of facial expression analysis, which represents the main focus of our research, since subtle

Point name	Symbol	Definition
Right outer eye corner	<i>A</i>	Point near the side of the face where the outer folds of the upper and lower right eyelid come together. This is not on the edge between the eyelid and the white of the eye but on the other side of the fold of the eyelids
Left outer eye corner	<i>A1</i>	Point near the side of the face where the outer folds of the upper and lower right eyelid come together. This is not on the edge between the eyelid and the white of the eye but on the other side of the fold of the eyelids
Right inner eye corner	<i>B</i>	Point near the nose where the upper and lower right eyelid come together, selected on the inside of the eye, where the tear gland begins
Left inner eye corner	<i>B1</i>	Point near the nose where the upper and lower left eyelid come together, selected on the inside of the eye, where the tear gland begins
Right inner eye brow	<i>D</i>	Point horizontally on the edge between the brows and the skin of the head between the brows, vertically halfway on the width of the eyebrow
Left inner eye brow	<i>D1</i>	Point horizontally on the edge between the brows and the skin of the head between the brows, vertically halfway on the width of the eyebrow
Right outer eye brow	<i>E</i>	Point vertically on the edge between the skin of the forehead and the eye brow. Horizontally approximately above the outer eye corner. If a distinct shape feature is present (like an arch or corner) select this point, otherwise select point right above the relevant outer eye corner
Left outer eye brow	<i>E1</i>	Point vertically on the edge between the skin of the forehead and the eye brow. Horizontally approximately above the outer eye corner. If a distinct shape feature is present (like an arch or corner) select this point, otherwise select point right above the relevant outer eye corner
Right upper eyelid	<i>F</i>	Point on the right upper eyelid, horizontally on the highest part of the eyelid, vertically immediately above the line where the eyelashes attach
Left upper eyelid	<i>F1</i>	Point on the left upper eyelid, horizontally on the highest part of the eyelid, vertically immediately above the line where the eyelashes attach
Right lower eyelid	<i>G</i>	Point on the right lower eyelid, horizontally on the lowest part of the eyelid, vertically immediately below the line where the eyelashes attach
Left lower eyelid	<i>G1</i>	Point on the left lower eyelid, horizontally on the lowest part of the eyelid, vertically immediately below the line where the eyelashes attach
Right nostril	<i>H</i>	Point on the right nostril, vertically halfway between the tip of the nose and where the nostril attaches to the face, horizontally in the middle of the width of the nostril
Left nostril	<i>H1</i>	Midpoint on the left nostril, halfway between the tip of the nose and where the nostril attaches to the face
Right mouth corner	<i>I</i>	Rightmost point where the skin parts to form the two lips ¹
Left mouth corner	<i>J</i>	Leftmost point where the skin parts to form the two lips
Upper lip	<i>M</i>	Point on the line where the red of the lower lip changes to the overall skin colour, horizontally in the middle of the philtrum
Lower lip	<i>L</i>	Point on the line where the red of the lower lip changes to the overall skin colour, halfway between the two mouth corners
Chin	<i>M</i>	Lowest point on the chin boss, before chin arches towards the neck
Nose	<i>N</i>	Tip of the nose: the point that is closest to the camera

Table 4.1: Definition of the 20 facial points used in our system.

changes in the facial feature appearance will be missed due to the errors in point localisation and subsequent tracking.

The goal of our fiducial facial point detector is to detect 20 fiducial facial points (illustrated in Fig. 4.1) plus the irises and the medial point of the mouth in the face region. In table 4.1 we give a definition of these points, which is needed for manual annotation of the points for training and evaluation of the facial point detection system. Left and right is from the subject's point of view. The description assumes a frontal upright pose.

These fiducial facial points are detected in four consecutive steps. First we find the face region in the input image. Next, the detected face region is roughly divided into 20 rectangular regions of interest (ROIs), each corresponding to one facial point to be detected. The employed method then

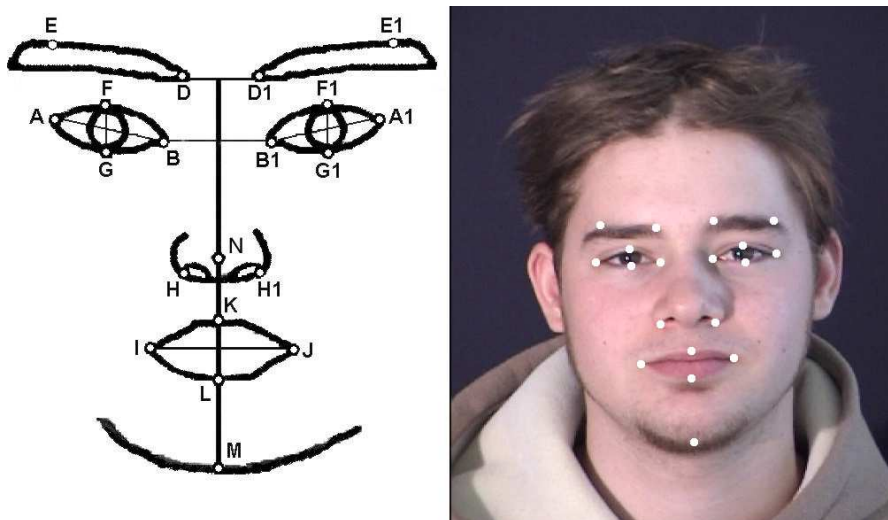


Figure 4.1: Fiducial facial point model (left) and fiducial facial points annotated on an image of a neutral face from the MMI Facial Expression Database

computes feature patches to detect points in the relevant ROI. These feature patches are constructed by concatenating for a fixed sized image patch both the gray level intensities and the Gabor filter bank responses into one large vector. We do not use histogramming to ensure that we capture the structure contained in the patch. We then use a GentleBoost classifier to select the patch at the coordinate inside an ROI that best represents the facial point we are looking for. In the following sections we will explain in detail how each step is performed. Figure 4.2 shows a graphical outline of the method.

4.1 Face detection

The first step in any fully automatic facial expression analysis method is face detection, i.e., identification of all regions in the scene that contain a human face. Numerous techniques for face detection in still images have been developed [Yang et al., 2002, Li and Jain, 2005]. The methods that had the greatest impact on the computer vision community (measured by, e.g, citations) include the following.

A multi-layer neural network was used by Rowley et al. [Rowley et al., 1998] to discern face from non-face patterns using the intensities and spatial relationships of pixels in face and non-face images. Sung and Poggio [Sung and Poggio, 1998] proposed a similar method, using a neural network to find a discriminant function based on distance measures. Moghaddam and Pentland developed a probabilistic visual learning method based on density estimation in a high-dimensional space using an eigenspace decomposition [Moghaddam and Pentland, 1997]. The method has been applied to face detection,

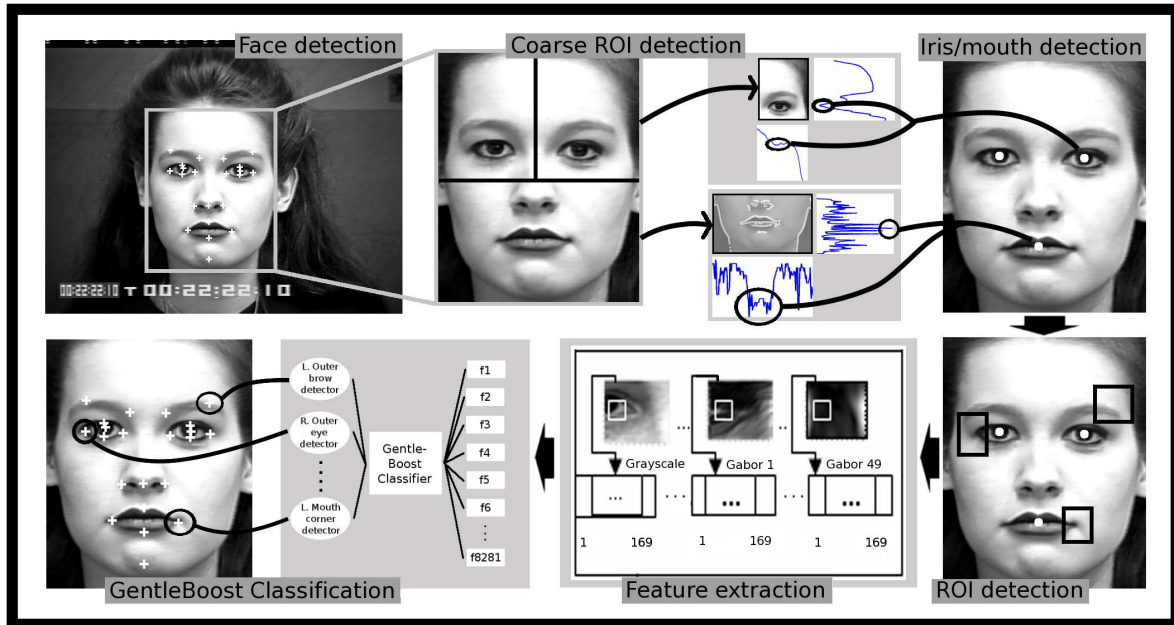


Figure 4.2: Outline of the fiducial facial point detection system.

coding and object recognition. Pentland et al. [Pentland et al., 1994] developed a real-time, view-based and modular eigenspace description technique for face recognition, where the face could be in various poses relative to the observer.

More recent work in face detection include Li et al. [Li et al., 2005], who learn multiview face subspaces using Independent Component Analysis (ICA). They showed that ICA, working on grey-scale images, was able to learn without supervision subspaces for various head poses in images, while Principal Component Analysis could not. Using ICA they could both detect faces with arbitrary head pose and could get a coarse head pose estimation. Hayashi and Hasegawa proposed a system that is able to detect faces from extremely low resolution images where the face patch is only 6x6 pixels big [Hayashi and Hasegawa, 2006]. Two detectors using different parameters for creating Haar-basis features were applied to the input images and their output was fused to a final classification result by a SVM.

Perhaps the most significant face detection method to date was proposed in 2001 by Viola and Jones [Viola and Jones, 2001]. The Viola-Jones face detector consists of a cascade of classifiers trained by AdaBoost. Each classifier operates on a Haar Basis representation of the input image. It uses integral histogram searching [Porikli, 2005] to enable a fast search on many locations and scales. While this

integral imaging process is essential to the speed of the detector, it restricts the detector to analysing rectangular regions². For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost. To detect the face in the first frame of an input image sequence scene we make use of a variant of this real-time face detector, as proposed in [Fasel et al., 2005].

The adapted version of the Viola-Jones face detector that we employ uses GentleBoost instead of AdaBoost. GentleBoost has been shown to be more accurate and converges faster than AdaBoost (see [Friedman et al., 2000] and a discussion in section 6.3.1). It uses a basic set of Haar basis filters, the responses of which are the feature values. In other words, each filter represents one feature. At each feature selection step (i.e., for every feature selected by GentleBoost), the proposed algorithm refines the feature originally proposed by GentleBoost. The algorithm creates a new set of image filters by placing the original filter and slightly modified versions of the filter at a two pixels distance in each direction and then computes the features returned by this image filter. The modified filters are created by scaling the original filter up to two pixels and reflecting each filter horizontally about the centre and superimposing it on the original. The modified filters are thus more complex variations of the original filters and so this process can be thought of as a single-generation genetic algorithm. Because we start of using only a relatively small number of filters and we only compute the responses of modified filters if its parent filter was found to be successful for the face detection task, it can be expected to be much faster and to achieve better results than searching the responses of all possible filters (features). Finally the employed version of the face detector uses a smart training procedure in which, after each single feature, the system can decide whether to test another feature or to stop and make a decision.

The face detector was evaluated on a set of 422 images with neutral expressions taken from the Cohn-Kanade face database [Kanade et al., 2000]. This set of images is representative for the type of imagery that we target in our systems: near-frontal views of faces recorded under normal lighting conditions. The recognition rate on this set was 100%.

4.2 ROI detection

The face detector returns a bounding box that delimits the face (i.e., the ‘face box’). We divide this face box into ROIs, one for each facial point we wish to detect. Because the locations of the ROIs

²This could be problematic when non-upright faces are present. An image pre-processing step would be needed that rotates the input image over a number of predefined angles to overcome this problem.

are defined using anatomic relations based on the locations of the irises and the medial point of the mouth, these points are localised first.

The iris and medial point of the mouth are detected as follows. First, we divide the face region horizontally in two: the upper face region containing the eyes and the lower face region containing the mouth. Since the utilised face detector is highly accurate, the face region contains all 20 characteristic facial points and little to no non-face regions (such as areas containing the hair, see fig. 4.2). It is therefore sufficient for the first step to roughly divide the face region horizontally in two halves. The upper face region is again divided vertically in two halves so that each eye can be localised separately.

The irises are localised in the segmented eye regions by sequentially applying the analysis of the row-wise sum of intensities (depicting the intensity differences between the successive rows) followed by the analysis of the column-wise sum of intensities (depicting the intensity differences between the successive columns). The minimum of the row intensities of the eye-region box corresponds to the y-coordinate of an iris, and the first minimum of the column-wise intensity sums of the eye-region box corresponds to the x-coordinate of an iris (see figure 4.2).

To describe the texture of the face, we intend to use Gabor filters, which we will describe in some detail in section 4.3. Although the Gabor filters are relatively insensitive to shift and scale variance, the grey-level features are not. The Gabor filters will also have a different response to a rotated input image. Therefore, registration of the input face image is mandatory. When the spatial coordinates of both irises are known, we are able to register the face by means of two parameters: the interocular angle α_I that the line connecting the irises makes with the horizontal image axis and the interocular distance D_I between the irises. The input face box is rotated using α_I and scaled by a factor $\frac{100}{D_I}$, so that the registered face is upright with the interocular distance $D'_I = 100$ pixels. From this point onward, distances measured in pixels assume that we have registered the face image.

The registration method we use here is based on two points and is thus only capable of coping with in-plane rotations (roll) and scale variations. It furthermore assumes that all points of the face lie on a plane. This means that our system will have great difficulty with faces that have a pitched or yawed pose with respect to the camera (i.e. out-of-plane rotations). If we imagine that there is large pitch (e.g. the person is looking down), the registration method will find that, as the positions of the eyes have only moved down a little, $\alpha_I \approx 0$ and D is the same as the upright face, too. Therefore, no registration will be carried out, while this would obviously be necessary. Similarly, if there would

have been significant pitch but no zoom, D would decrease as the projection of the eyes on the camera image plane places the eyes closer together. This would cause an unwanted transformation, as the decreased D invokes a transformation to cancel the (incorrectly perceived) zoom. In fact, what we would like is a general purpose 3D-model of a face, that uses only 2D input information to compute a 3D transformation of the input image. However, the implementation of 2D to 3D techniques goes beyond the scope of this thesis and is considered future work.

To locate the medial point of the mouth, we first apply an edge detector to the input image. Such an edge representation of the lower face region is illustrated in Fig. 4.2. Analysis of the histogram of the edge representation gives us the position of the medial point of the mouth. The centre of the widest peak will define the vertical position. By selecting the widest peak of the histogram, the possibility of detecting the nostrils instead of the mouth is avoided. The horizontal position of the medial mouth point is then defined as the position that corresponds to the highest peak of the histogram of the line defined by the mouth's vertical position (see figure 4.2).

Subsequently, we use the detected positions of the irises and the medial point of the mouth to divide the face into 20 regions, based on rules defined by the anatomy of the human face. Each of the points to be localised is within one ROI. An example of ROIs extracted from the face region for points B, I, and J (see figure 4.1), is depicted in Fig. 4.2. The method is capable of detecting facial points in images of neutral frontal-view faces. Because of this neutrality and frontal-view constraints, we tested the system on all the first frames of the Cohn-Kanade database, a total set of 422 images. On this set we achieved a detection rate of 100% (i.e., all the segmented ROIs were correctly positioned to contain their respective facial feature point).

4.3 Gabor Feature Extraction

The proposed facial feature point detection method uses individual feature patches to detect points in the relevant ROI. The feature models are defined as GentleBoost templates learned from both the gray level intensity and Gabor filter response representation of square 13x13 pixels image patches centred around a point.

Recent work [Donato et al., 1999] has shown that a Gabor filter approach for local feature extraction outperforms PCA (Principal Component Analysis), FLD (Fisher's Linear Discriminant) and

LFA (Local Feature Analysis). The essence of the success of Gabor filters is that they remove most of the differences between images that are caused by variation in lighting and contrast or by small translation and deformation transformations [Osadchy et al., 2007]. Gabor filters seem to be a good approximation of the sensitivity profiles of neurons found in the visual cortex of higher vertebrates [Jones and Palmer, 1987]. There is evidence that those cells come in pairs with even and odd symmetry [Field, 1978, Burr et al., 1989], similar to the real and imaginary part of Gabor filters. Therefore it seems that Gabor filters are a good approach for computer vision problems that simulate human vision tasks.

A 2D Gabor filter $\psi(x, y)$ can be defined as:

$$\psi(x, y) = \frac{\alpha\beta}{\pi} e^{-(\alpha^2 x'^2 + \beta^2 y'^2)} e^{(2\pi j f_0 x')} \quad (4.1)$$

with

$$\begin{aligned} x' &= x \cos \theta + y \sin \theta \\ y' &= -x \sin \theta + y \cos \theta \end{aligned} \quad (4.2)$$

where f_0 is the central frequency of a sinusoidal plane wave, j the imaginary unit, θ is the radial rotation of the Gaussian envelope and the plane wave, and α and β are the parameters for scaling two axes of the elliptic Gaussian envelope. Note that equation (4.1) ensures that the orientation of the Gaussian envelope and the orientation of the sinusoidal carrier f_0 are the same, which is one of the characteristics of complex cells of the mammals' visual cortex [Jones and Palmer, 1987]. The Gabor function is actually a sinusoidal plane wave (second part of equation (4.1)) modulated by a Gaussian shaped function (first part of equation (4.1)). Its 2D Fourier transform is:

$$\psi(u, v) = \exp \left(-\pi^2 \left\{ \frac{(u \cos \theta + v \sin \theta)^2}{\alpha^2} + \frac{(u \sin \theta - v \cos \theta)^2}{\beta^2} \right\} \right) \quad (4.3)$$

By fixing the ratio of the wave frequency and the sharpness of the Gaussian we maintain a constant number of waves in the spatial filter (4.1). The ratios which are known to satisfy this property for the complex cells in the human visual cortex are [Jones and Palmer, 1987]:

$$\gamma = \frac{f_0}{\alpha} = \frac{1}{\sqrt{\pi \cdot 0.9025}} \quad (4.4)$$

$$\eta = \frac{f_0}{\beta} = \frac{1}{\sqrt{\pi \cdot 0.58695}} \quad (4.5)$$

Thus, by using 4.4 and 4.5 in equations 4.1 and 4.3 a normalised filter can be presented in spatial domain as:

$$\psi(x, y) = \frac{f_0^2}{\pi \gamma \eta} \exp \left\{ - \left(\frac{f_0^2}{\gamma^2} x'^2 + \frac{f_0^2}{\eta^2} y'^2 \right) \right\} \exp(2\pi j f_0 x') \quad (4.6)$$

and in frequency domain as:

$$\Psi(u, v) = \exp \left(- \frac{\pi^2}{f_0^2} \{ \gamma (u' - f_0) + \eta^2 v'^2 \} \right) \quad (4.7)$$

where

$$\begin{aligned} u' &= u \cos \theta + v \sin \theta \\ v' &= u \sin \theta + v \cos \theta \end{aligned} \quad (4.8)$$

Thus, in the frequency domain the filter is an oriented Gaussian with orientation θ centred at frequency f_0 . A Gabor filter formulated in this way has a zero frequency response value (DC-response) close to 0 and is the same for all central frequencies. This ensures that the method is insensitive to global illumination variations.

Several Gabor filters can be combined to form a filter bank. The filter bank is usually composed of filters in various orientations and frequencies, with equal orientation spacing and octave frequency spacing, while the relative widths of Gaussian envelope γ and η stay constant. In the frequency domain the Gabor filter must obey the Nyquist rule, which means that $(\forall \theta) f_0 \leq 0.5$.

We extract a feature vector for a certain location in an image from the 13x13 image patch centred on that point. As features we use the gray scale image information of that patch plus the responses from a bank of 48 Gabor filters at 8 orientations and 6 spatial frequencies (2:12 pixels/cycle at 1/2 octave steps). We thus attain the set F_p consisting of $169 \times 49 = 8281$ features to represent a point.

4.4 Facial Point Detection using GentleBoost

We use a GentleBoost classifier to identify the location of a target point within its ROI. Because of the high dimensionality of our data, boosting techniques are very suitable classifiers, as they implicitly perform feature selection in the training stage. The small number of features actually employed by the trained GentleBoost classifier ensures fast detection of the characteristic facial points and improves generalisability. We have chosen GentleBoost instead of AdaBoost as it converges faster than AdaBoost, and performs better for object detection problems in terms of accuracy [Torralba et al., 2004]. It is simple to implement, it is numerically robust and it has been shown experimentally to outperform other boosting variants for the face detection task [Lienhart et al., 2003] (with respect to detection accuracy).

The performance of boosting methods on data which are generated by classes that have a significant overlap, in other words, classification problem where even the Bayes error is significantly high is discussed in [Freund and Schapire, 2000]. For this case, GentleBoost performs better than AdaBoost since AdaBoost over-emphasises the atypical examples which eventually results in inferior rules. As explained in [Freund and Schapire, 2000], the reason for this might be that GentleBoost gives less emphasis to misclassified examples since the increase in the weight of the example is linear in the negative margin, rather than exponential.

For the AU recognition approach described in chapter 6, we use GentleBoost as a feature selector preceding a more powerful SVM classifier. We have not applied this approach to the problem of characteristic facial point detection, as the good results using GentleBoost only (see section 4.5) indicated that this more complicated and computationally expensive approach was not needed.

The outline of the GentleBoost algorithm is as follows. As weak classifiers we use a simple linear function with variables a and b and threshold ζ to approximate the labels y given a one-dimensional set of features x :

$$y = a(x > \zeta) + b \quad (4.9)$$

At each boosting round, a weak classifier is trained for every feature. These are our candidate weak classifiers. We find for each feature j the optimal values for a_j , ζ_j and b_j , measured by weighted least-squared error. In our case, the feature is either a gray level value or a Gabor filter response. The

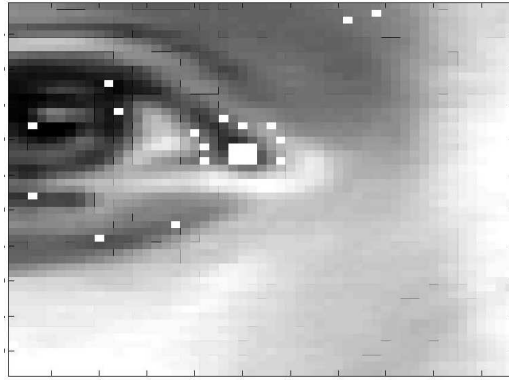


Figure 4.3: Positive and negative examples for training the classifier for the right inner eye corner. The big white square represents the 9 positive examples. Eight negative examples have been randomly picked near the positive examples and another 8 are randomly chosen from the remainder of the region of interest.

fitting of the regression function is done for every feature $1 \leq j \leq |F_p|$ for all training examples by minimising the weighted error η_j :

$$\eta_j = \operatorname{argmin}_{a,b,\zeta} \frac{\sum_i (w_i |y_i - (a_j (x_{i,j} > \zeta_j) + b_j)|^2)}{\sum_i w_i} \quad (4.10)$$

where i is the i -th training example, $x_{i,j}$ the value of the j -th feature of training example i and y_i is its class label, $y_i \in \{0, 1\}$. Here we use equation 4.9 to determine the class of instance i . We apply equation 4.10 to all features j , and select the feature with lowest error η_j , denoted by J . The weak classifier trained for this round is then defined by the corresponding linear function parameters a_J , b_J and ζ_J :

$$f_m(x_i) = a_J (x_{i,J} > \theta_J) + b_J \quad (4.11)$$

where J is the feature chosen at round m . The next step is to update the strong classifier $F(x_i, m-1)$ that is an ensemble of the $m-1$ previously selected weak classifiers $f(x)$. The output of the strong classifier after m rounds is found as follows:

$$F(x_i, m) = F(x_i, m-1) + f_m(x_i) \quad (4.12)$$

and the updated weights for each training example after round m are found as:

$$w_i \leftarrow w_i e^{-y_i f_m(x_i)} \quad (4.13)$$

The sequential selection of a new weak classifier, calculation of the strong classifier output and updating of the weights continues until the output of the strong classifier converges with the class labels of the training data, or some other stop condition has been met. Finally, the weights of the samples should be re-normalised. Now, for a new testing example x' the output of the GentleBoost classifier can be calculated as:

$$F(x') = \sum_{m=1}^M f_m(x') \quad (4.14)$$

where M is the number of the most relevant features, i.e., the value of m at the time that the GentleBoost classifier converged. If this were a binary classification problem, the class C of instance x is then found as $C = \text{sign}(F)$. However, for the problem of facial point detection we will keep working with the real values $F(x)$.

In the training phase, GentleBoost feature templates are learned using a representative set of positive and negative examples. As positive examples for a facial point, we used the features derived from a block of 9 points centred on the (manually selected) true point. For each facial point we used two sets of negative examples. The first set contains 8 randomly chosen image patches that are 2 to 4 pixels separated from the 9 positive sample points. The second set contains 8 image patches randomly displaced in the relevant ROI (Fig. 4.3), with a minimum distance of 4 pixels to the positive samples. Thus, for each target point, we have 9 positive and 16 negative examples, resulting in a set with dimensions 25×8281 representing the training data for each point in a training image. Even though each feature can be computed very efficiently, computing the complete set is computationally expensive. Adding the fact that the feature representation is highly redundant, the choice of boosting techniques seems justified.

Eventually, in the testing phase, each ROI is filtered first by the same set of Gabor filters used in the training phase. Then, for detection of the relevant characteristic facial point an input 13x13 pixels sliding window is slid pixel by pixel across the ROI. For each position of the sliding window, the GentleBoost classifier returns a response which is a measure of similarity between the feature vector retrieved at the current position of the sliding window and the learned feature point model. After scanning the entire ROI, the position with the highest response (i.e., with the largest positive value



Figure 4.4: Examples of accurately detected facial points.

of $F(x)$ is chosen as the location of the feature point in question.

4.5 Evaluation

To evaluate the performance of the point detector, we measured the distance between the automatically detected points and the manually annotated points. We consider a point to be located correctly if the distance e_p between the automatically detected point and the manually annotated point is less than 10% of the IOD, i.e. $e_p < 0.1D_I$. The system was evaluated both on the Cohn-Kanade database and on the MMI database. When tested on 300 images taken from the Cohn-Kanade database in a three-fold cross-validation evaluation, the system achieved an average 93% recognition rate for all points on all folds. Table 4.2 shows the results for every facial point separately. When trained on all 300 images taken from the Cohn-Kanade database and tested on 244 images taken from the MMIDatabase, an average 92.2% classification rate was achieved. Table 4.3 shows the results for the test on the MMI Database. In this table, we have also listed the average error e_p per facial point, measured in units of D_I . Figure 4.4 shows some typical point detection results.

	Detected point	Cl. Rate		Detected point	Cl. Rate
A	Right eye outer corner	0.92	G	Right eye lower eyelid	0.95
A1	Left eye outer corner	0.96	G1	Left eye lower eyelid	0.99
B	Right eye inner corner	0.96	H	Right nostril	0.98
B1	Left eye inner corner	0.99	H1	Left nostril	0.97
D	Right brow inner corner	0.96	I	Right mouth corner	0.97
D1	Left brow inner corner	0.95	J	Right mouth corner	0.91
E	Right brow outer corner	0.96	K	Mouth top	0.93
E1	Left brow outer corner	0.90	L	Mouth bottom	0.80
F	Right eye upper eyelid	0.91	M	Chin	0.90
F1	Left eye upper eyelid	0.83	N	Tip of nose	0.98
	Average for all points:	0.93			

Table 4.2: Characteristic Facial Point detection results for 300 samples from the Cohn-Kanade database

	Cl. rate	e_p		Cl. rate	e_p
A	0.784	0.079	G	0.982	0.026
A1	0.976	0.045	G1	0.982	0.018
B	0.976	0.024	H	0.976	0.030
B1	0.952	0.035	H1	0.976	0.036
D	0.569	0.093	I	0.904	0.068
D1	0.802	0.075	J	0.928	0.053
E	0.928	0.035	K	0.964	0.037
E1	0.958	0.035	L	0.952	0.042
F	0.982	0.037	M	0.904	0.052
F1	0.982	0.043	N	0.952	0.058
Avg	0.922	0.046			

Table 4.3: Characteristic Facial Point detection results for 244 samples from the MMI-Facial Expression Database. e_p is the average relative error, measured in units of Interocular Distance D_I

Chapter 5

Feature extraction

Our AU analysis system uses spatio-temporal facial geometry-based features computed from tracked facial feature points. In this chapter we will explain how the previously detected facial points (see chapter 4) are tracked and how the features that are used for AU activation detection and AU temporal phase detection are extracted from the tracking data.

5.1 Facial feature point tracking

After localisation of the 20 characteristic facial points in the first frame of an input face video (see chapter 4), the points are tracked in all subsequent frames. Standard optical flow techniques are commonly used for facial point tracking in facial expression analysis. For example, to achieve facial point tracking, Tian et al. [Tian et al., 2001] and Cohn et al. [Cohn et al., 2004b] use the standard Lucas-Kanade optical flow algorithm [Lucas and Kanade, 1981], and Xiao et al. [Xiao et al., 2003] use an “inverse compositional” extension to the Lucas-Kanade algorithm to realize fitting of 2D and combined 2D+3D Active Appearance Models to images of faces. To omit the limitations inherent in optical flow methods, such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and changes in illumination, few researchers used sequential state estimation techniques to track facial points in image sequences.

Both, Zhang and Ji [Zhang and Ji, 2005] and Gu and Ji [Gu and Ji, 2005] used facial point tracking based on a Kalman filtering scheme, which is the traditional tool for solving sequential state problems. The derivation of the Kalman filter is based on a state-space model governed by two assumptions

[Kalman, 1960]: (i) linearity of the model and (ii) Gaussian distribution of both the dynamic noise in the process equation and the measurement noise in the measurement equation. Under these assumptions, derivation of the Kalman filter leads to an algorithm that propagates the mean vector and covariance matrix of the state estimation error in an iterative manner and is optimal in the Bayesian setting.

To deal with the state estimation in nonlinear dynamical systems, the extended Kalman filter has been proposed, which is derived through linearisation of the state-space model. However, many of the state estimation problems, including human facial expression analysis, are nonlinear and quite often non-Gaussian too. Thus, if the face undergoes a sudden or rapid movement, the prediction of features positions from Kalman filtering will be significantly off. To handle this problem, Zhang and Ji [Zhang and Ji, 2005] and Gu and Ji [Gu and Ji, 2005] used the information about the IR-camera-detected pupil location together with the output of Kalman filtering to predict facial features positions in the next frame of an input face video. To overcome these limitations of the classical Kalman filter and its extended form in general, particle filters have been proposed. In this section we will use Particle Filtering with Factorised Likelihoods (PFFL, [Patras and Pantic, 2004]). We will also give a basic introduction to particle filtering. For a detailed overview of the various facets of particle filters, see [Andrieu et al., 2004].

In recent years, particle filtering has been the dominant paradigm for tracking the state α of a temporal event given a set of noisy observations $Y = \{y^0 \dots, y^{t-1}, y^t\}$ up to the current time instant t . In our case, the state α is a description of the locations of a set of facial fiducial points while the observations in the set Y are the frames of an input face video.

The main idea behind particle filtering is to maintain a set of probable solutions that are an efficient representation of the conditional probability $p(\alpha|Y)$. By maintaining a set of solutions instead of a single estimate (as is done by Kalman filtering, for example), particle filtering is able to track multimodal conditional probabilities $p(\alpha|Y)$, and it is therefore robust to missing and inaccurate data and particularly attractive for estimation and prediction in nonlinear systems and systems where the probability distribution cannot be modeled by a Gaussian distribution.

Particle Filtering with Factorised Likelihoods is an extension to the Auxiliary Particle Filtering theory introduced by Pitt and Shephard [Pitt and Shephard, 1999], which itself is an extension to the classic Condensation algorithm [Isard and Blake, 1998]. As opposed to its predecessors, PFFL does not track

all points independently but tracks ‘constellations’ of facial points instead. Providing the tracker with *a priori* information about the probability of where the elements of these constellations are relative to each other allows the PFFL tracker to give more accurate predictions of the positions of the points. Fig. 5.1 shows typical results of the tracker on images from the MMI Facial Expression Database and the Cohn-Kanade Database. In the figure the top three rows are from the MMI database and the bottom two rows from the Cohn-Kanade database. All rows start with a neutral expression. The MMI database rows subsequently show one frame where the action is growing stronger (onset-phase), the third column shows the peak of the expression (apex-phase) and the fourth and fifth columns show frames where the action is diminishing again (offset phase). The rows for the Cohn-Kanade database show the onset-phase in the second through to the fourth column and the apex phase in the fifth and last column.

5.1.1 Condensation algorithm

The main idea of particle filtering is to maintain a particle based representation of the *a posteriori* probability $p(\alpha | Y)$ of the state α given all the observations Y up to the current time instance. This means that the distribution $p(\alpha | Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k , then it is as if s_k was drawn from $p(\alpha | Y)$. In the particle filtering framework our knowledge about the *a posteriori* probability is updated in a recursive way. Suppose that at a previous time instance we have a particle based representation of the density $p(\alpha^- | Y^-)$, that is, we have a collection of K particles and their corresponding weights (i.e. $\{(s_k^-, \pi_k^-)\}$). Then, the Condensation Particle Filtering can be summarised as follows:

1. Draw K particles s_k^- from the probability density that is represented by the collection $\{(s_k^-, \pi_k^-)\}$.
2. Propagate each particle s_k^- with the transition probability $p(\alpha | \alpha^-)$ in order to arrive at a collection of K particles s_k .
3. Compute the weights π_k for each particle as follows,

$$\pi_k = p(y | s_k) \tag{5.1}$$

Then normalise so that $\sum_k \pi_k = 1$.

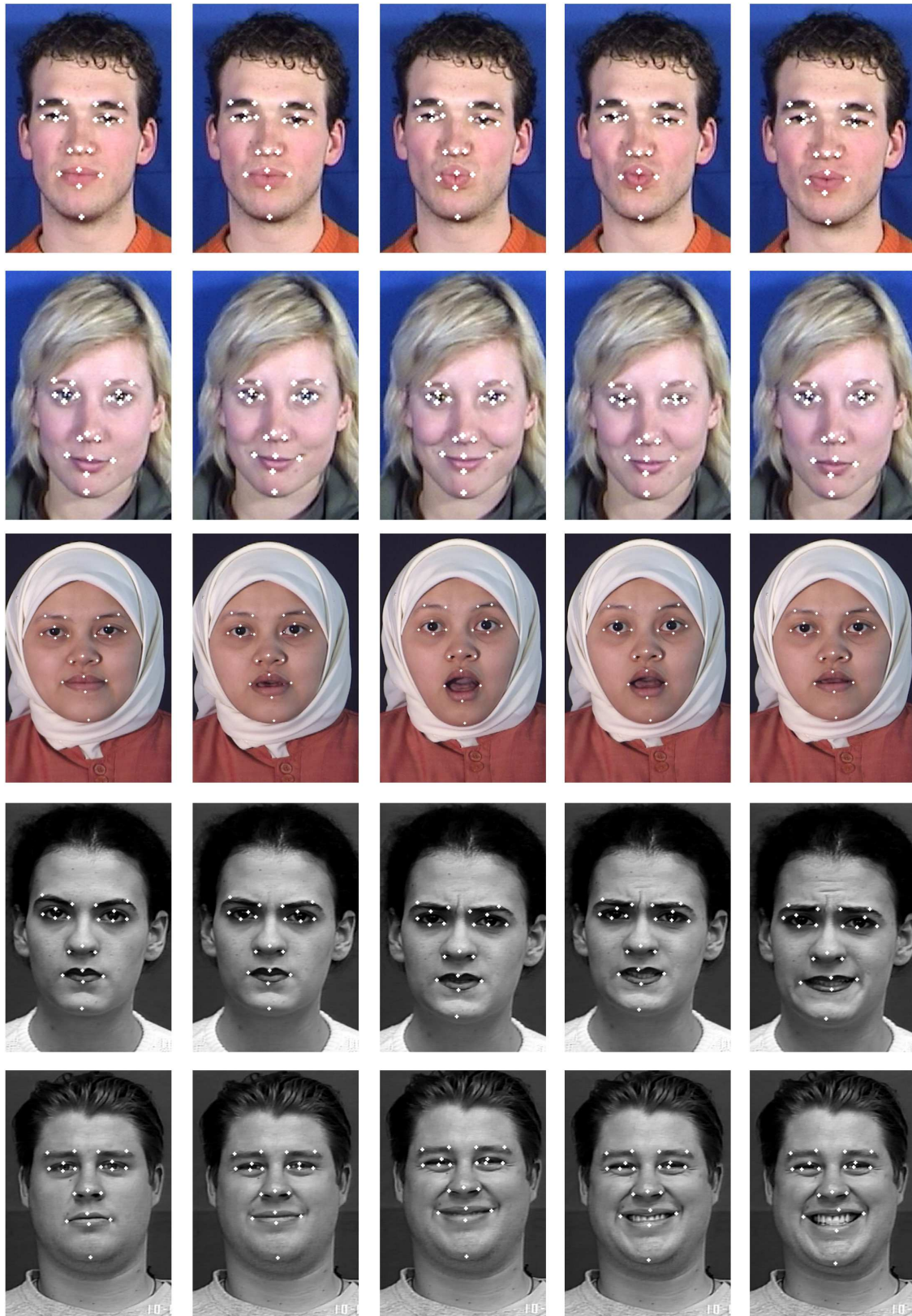


Figure 5.1: Results from the fiducial facial point tracker, which uses Particle Filtering with Factorized Likelihoods. 20 fiducial facial points were tracked in 332 sequences of up to 200 frames taken from the MMI Database and the Cohn-Kanade database. The top three rows are from the MMI database and the bottom two rows from the Cohn-Kanade database. All rows start with a neutral expression. The

This results in a collection of K particles and their corresponding weights (i.e. $\{(s_k, \pi_k)\}$ which is an approximation of the density $p(\alpha|Y)$.

5.1.2 Particle filtering with Factorised likelihoods

The Condensation algorithm has three major drawbacks. The first drawback is that a large amount of particles that result from sampling from the proposal density $p(\alpha|Y^-)$ might be wasted because they are propagated into areas with small likelihood. The second problem is that the scheme ignores the fact that while a particle with an N dimensional state $s_k = \langle s_{k1}, s_{k2}, \dots, s_{kN} \rangle$ might have a low likelihood, it can easily happen that parts of it might be close to the correct solution. Finally, the third problem is that the estimation of the particle weights does not take into account the interdependencies between the different parts of the state α .

Particle filtering with Factorised likelihoods [Patras and Pantic, 2004] attempts to solve these problems in one step, given the case that the likelihood can be Factorised, that is, in the case that $p(y|\alpha) = \prod_i p(y|\alpha_i)$. It uses as proposal distribution $g(\alpha)$ the product of the posteriors of each α_i given the observations, that is $g(\alpha) = \prod_i p(\alpha_i|y)$, from which we draw samples s_k . These samples are then assigned weights π_k , using the same proposal distribution. We now find π_k and s_k as follows:

1. Propagate all particles s_k^- via the transition probability $p(\alpha_i|\alpha^-)$ in order to arrive at a collection of K sub-particles μ_{ik} . Note, that while s_k^- has the dimensionality of the state space, the μ_{ik} have the dimensionality of the partition i .
2. Evaluate the likelihood associated with each sub-particle μ_{ik} , that is let $\lambda_{ik} = p(y|\mu_{ik})$.
3. Draw K particles s_k^- from the probability density that is represented by the collection $\{(s_k^-, \lambda_{ik}\pi_k^-)\}$.
4. Propagate each particle s_k^- with the transition probability $p(\alpha_i|\alpha^-)$ in order to arrive at a collection of K sub-particles s_{ik} . Note, that s_{ik} has the dimensionality of the partition i .
5. Assign a weight π_{ik} to each sub particle as follows, $w_{ik} = \frac{p(y|s_{ik})}{\lambda_{ik}}$, $\pi_{ik} = \frac{w_{ik}}{\sum_j w_{ij}}$. With this procedure, we have a particle-based representation for each of the N posteriors $p(\alpha_i | y)$. That is, we have N collections $(s_{ik},) \pi_{ik}$, one for each i .
6. Sample K particles from the proposal function $g(\alpha) = \prod_i p(\alpha_i | Y)$. This is approximately equivalent to constructing each particle $s_k = \langle s_{k1} \dots s_{ki} \dots s_{kN} \rangle$ by sampling independently each s_{ik} from $p(\alpha_i | Y)$.

7. Assign weights π_k to each particle as follows:

$$\pi_k = \frac{p(s_k|Y^-)}{\prod_i p(s_{ik}|Y^-)} \quad (5.2)$$

8. Finally, the weights are normalised to sum up to one.

With this, we end up with a collection $\{(s_k, \pi_k)\}$ that is a particle-based representation of $p(\alpha|Y)$. Note that at the numerator of eq. 5.2 the inter-dependencies between the different sub-particles are taken into consideration. On the contrary, at the denominator, the different sub-particles are considered independent. In other words, the re-weighting process of eq. 5.2 favours particles for which the joint is higher than the product of the marginals.

5.1.3 Rigid and morphological observation models

In steps 2 and 5 of the PFFL algorithm described in section 5.1.2, the likelihood and the weight of each sub-particle are determined by applying an observation model. We use two different models. Both models are robust, invariant to lighting conditions, colour-based observation models for template-based tracking. The first model is suitable for tracking the motion of image patches around facial features that do not change their appearance significantly. The second model allows for minor morphological transformations of the image patch, but is computationally more expensive. The models are initialised in the first frame of an image sequence when a set of $n = 20$ windows are centred around the characteristic facial points detected by the point detector described in chapter 4 and that will be tracked through the remainder of the image sequence. Let us denote with \mathbf{o}_i the template feature vector, which contains the RGB colour information at window i in frame 1.

We need to define $p(y|\alpha_i)$. Let us denote with $y(\alpha_i)$ the template feature vector that contains the RGB colour information at the image patch around point α_i . We use a colour-based difference between the vectors \mathbf{o}_i and $y(\alpha_i)$ that is invariant to global changes in the intensity as follows:

$$c(\mathbf{o}_i, y(\alpha_i)) = \left(\frac{\mathbf{o}_i}{E\{\mathbf{o}_i\}} - \frac{y(\alpha_i)}{E\{y(\alpha_i)\}} \right) \quad (5.3)$$

where $E\{\mathbf{x}\}$ is the (scalar) average taken over the average intensities of all colour channels on a colour template \mathbf{x} . It is easy to show that the colour difference vector $c(\mathbf{o}_i, y(\alpha_i))$ is invariant to global

changes in the light intensity¹. Finally, we define the scalar colour distance using a robust function ρ . Let us denote with j the index of the colour difference vector j , that is $c_j(\mathbf{o}_i, y(\alpha_i))$, the difference in a specific colour channel at a specific pixel. The scalar colour distance is then defined as:

$$d_c(\mathbf{o}_i, y(\alpha_i)) = E_j |\rho(c_j(\mathbf{o}_i, y(\alpha_i)))| \quad (5.4)$$

where the robust function ρ that has been used in our experiments is the L_1 norm.

The second model allows for non-rigid deformations of the initial template, as mentioned above. Let us denote this unknown transformation by $\phi : N^2 \rightarrow N^2$, a transformation that gives the correspondence between the pixel coordinates of the colour template \mathbf{o}_i and the image patch $y(\alpha_i)$. Then, let us denote with $y(\alpha_i, \phi)$ the template that results after the nonrigid transformation ϕ is applied to the image patch $y(\alpha_i)$. The distance metric d_m between the initial template \mathbf{o}_i and $y(\alpha_i)$ contains two terms: the first term, $d_c(\mathbf{o}_i, y(\alpha_i, \phi))$, is similar to the distance measure $d_c(\mathbf{o}_i, y(\alpha_i))$ for the rigid observation model, only now we take the minimum colour distance over all possible deformations ϕ . The second term, $d_s(\phi)$, is a measure of the shape deformation that is introduced by the transformation ϕ . The morphology term $d_s(\phi)$ is defined as the average Euclidean distance over the pixel based displacements, that is

$$d_s(\phi) = E_i (\|i - \phi(i)\|_2) \quad (5.5)$$

where $\|x\|_2$ is defined as the L_2 norm of x and, with a slight abuse of notation, i denotes pixel coordinates. The distance measure is thus defined as:

$$d_m(\mathbf{o}_i, y(\alpha_i)) = \min_{\phi} (d_c(\mathbf{o}_i, y(\alpha_i, \phi)) + \lambda d_s(\phi)) \quad (5.6)$$

where the first term is used to penalise large colour-based distances, the second term is used to penalise large shape deformations and the parameter λ controls the balance between the two terms. Finally, the observation likelihood is calculated as follows:

$$p(y|\alpha_i, \mathbf{o}_i) = \frac{1}{z} \exp\left(\frac{-d_o(y(\alpha_i), \mathbf{o}_i)}{\sigma_i}\right) \quad (5.7)$$

¹Note that c contains the colour differences over all colour channels(R, G, B).

where σ_i is a scaling parameter for template i and $d_o(\mathbf{x}, \mathbf{y})$ is either the distance measure d_c defined in (5.4) or the distance measure d_m defined in (5.6) depending on which observation model is applied. The term z is a normalisation term, which in the particle filtering framework can be ignored, since the weights of the particles are renormalised at the end of each iteration to sum up to one.

The morphological model is used for points that strongly change their shape, in our case the mouth corners and eye corners. The rigid model is used for all other points. It is sufficiently accurate for representation of the target points, which do not significantly change in appearance during facial expressions, and it is cheaper to compute than the morphological model.

5.1.4 Transition model and priors

Once the observation model is defined, we need to model the transition probability density function that is used to generate a new set of particles given the current set and to specify the scheme for reweighting the particles by eq. (5.2).

The transition probability $p(\alpha_i|\alpha^-)$ models our knowledge of the dynamics of the features, that is, it models our knowledge of the feature's position $\alpha = \langle \alpha_i, \dots, \alpha_n \rangle$ in the current frame given its position α^- in the previous frame. To avoid sampling from a complicated distribution, we assume that $p(\alpha_i|\alpha^-) = p(\alpha_i|\alpha_i^-)$, i.e. we assume that each sub-particle can be propagated independently of other sub-particles. We use a simple zero order motion model with Gaussian noise, that is,

$$p(\alpha_i|\alpha^-) = \alpha_i^- + N(0, \sigma) \quad (5.8)$$

The reweighting scheme employed by the tracker uses training data to learn the *a priori* interdependencies between the positions of the facial microfeatures (i.e., facial characteristic points). In the utilised modelling, the fraction in eq. (5.2) is approximated by a prior on the relative positions of the facial points. This can be illustrated as shown in the graphical model of Fig. 5.2, where the correlations between various sub-particles — in our case facial points — are depicted by the dashed lines. This is evaluated with Parzen density estimation:

$$\pi_k = \sum_j \Xi[d(q(s_k), h_j), \sigma_p] \quad (5.9)$$

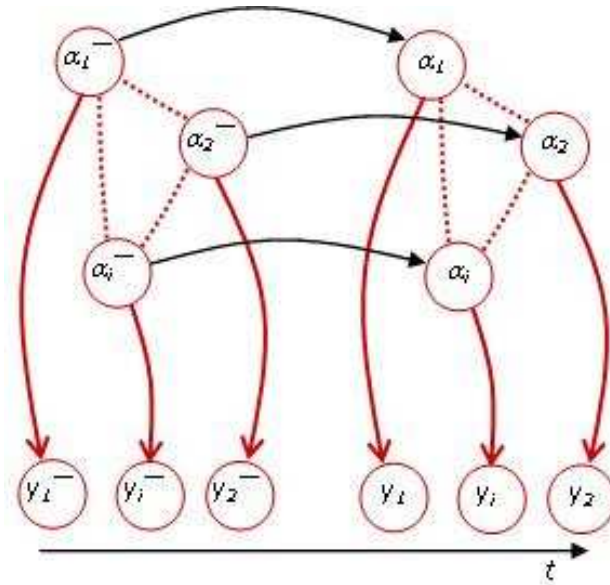


Figure 5.2: The assumed transition model. α_i and α_i^- are the current and the previous state of a facial point i and y_i and y_i^- are the current and the previous observations. Each facial point's state can be represented as a simple Markov chain in temporal domain whilst in each time instant different facial points may be related. The dashed lines represent those interdependencies.

where the collection $\{h_j\}$ is the collection of training data, $q(s_k)$ is an affine transformation function that registers the data from the current face (i.e., the particle s_k) to the training data (i.e., to the collection h_j), $d(a, b)$ is the Euclidian distance function between the registered particle a and a training datum b and $\Xi(x, y)$ is the Parzen density kernel function, in our case a Gaussian kernel with standard deviation σ_p .

As the collection of training data $\{h_j\}$, we used in our study sets of (semi-)automatically annotated data containing the coordinates of facial characteristic points belonging to four facial components: eyebrows, eyes, nose-chin and mouth. The underlying assumption is that correlations between the points belonging to the same facial components are more important for facial expression recognition than correlations between the points belonging to different facial components. This is consistent with psychological studies that suggest that: a) the brain processes facial expressions locally/analytically rather than holistically whilst it identifies faces holistically [Bassili, 1978], and b) dynamic cues (e.g. facial expressions) are computed separately from static cues (e.g. facial proportions) [Humphreys et al., 1993].

This dataset consists of 66 image sequences of 3 persons (one of which was female) showing the 23 AUs that the system presented in this paper is able to recognise. The utilised sequences are from the MMI facial expression database (posed expressions) [Pantic et al., 2005b] and they have not been used to train and test the performance of the other parts of the AU analysis system.

Each of these image sequences is annotated (semi-) automatically in two steps. First, the facial point detector described in chapter 4 is used to initialise 20 facial characteristic points in the first frame. Then, these points were tracked in the rest of the sequence by means of the Auxiliary Particle Filtering algorithm as explained in [Pitt and Shephard, 1999]. The positions of the automatically detected points were inspected by a human and, if necessary, corrected. Similarly, the obtained tracking data were inspected by a human and, if necessary, repeated using slightly different initialisation parameters to remove any tracking errors.

Finally, the data was registered using exactly the same procedure that is applied to register the data used for testing and training the system. In our experiments, this registered data forms the collection of training data h_j in eq. (5.9). Consequently, in our study, the function $q(s_k)$ is the identity transformation.

The PFFL tracker returns for every image sequence a point matrix P with dimensions $l \times 20 \times 2$, where l is the number of frames of the input image sequence.

5.2 Registration of tracking data

There are two registration problems that we have to address in order to be able to compare tracking data obtained for different face videos. The first issue is that we need to be able to differentiate between movement of the facial points due to motion of the head with respect to the camera (rigid head motion) and movement of facial points due to facial expressions. We must transform the tracking data of a video such that all rigid head motion (both translation and rotation) is cancelled. This is called intra-sequence registration, as this registration is done with respect to the position of the points found in the first frame of the input image sequence.

The other registration process addresses the problem that different people have different faces, with different dimensions, and that the face might be positioned each time differently in the first frame. We use a pre-defined, so-called *normal* face with respect to which all other faces are registered. This way, the facial points found in the first frame of a neutral face of every image sequence will be comparably positioned (i.e. on a fixed set of points in the point space). As this registration process is done between different image sequences, we call it inter-sequence registration.

5.2.1 Intra-sequence registration

Because in the intra-sequence registration we wish to separate point motion due to rigid head motion from point motion due to facial expression, we use the tracking data of stable facial points only. Stable facial points are facial points that do not move due to facial expressions, they will only move due to rigid head motion. From our 20 facial points (see Fig. 4.1), the inner and outer eye points (the points A , $A1$, B , and $B1$) are stable. Another stable point is the point just below the nose, on the philtrum. However, we cannot track this point due to frequent self occlusion by the nose. Therefore we instead use the mean of the points H and $H1$ to approximate the position of this fifth stable point. We denote this set of facial points by P_S . Now we can apply a simple transformation $T_A(P_S, t)$, an affine transformation without shearing, computed for each frame t by comparing the positions of the five stable points P_S at time t with their position at time $t = 1$, the first frame.

5.2.2 Inter-sequence registration

Different people have different faces. Some are long and square corners, others are small and round. The location of the face in the scene may vary from recording to recording, too. To overcome the related problems, all image sequences are registered with respect to a predefined face shape and location. This inter-registration process is also carried out by an affine transformation. Under the assumption that all image sequences begin with a neutral facial expression, for each new image sequence an affine transform $T_E(P)$ is computed by comparing the position of the facial points in the first (neutral) frame with these positions of the points in the predefined *normal* face². This transform is the same for all frames in a sequence, as it only takes into account the shape of the face and the initial position of the face. Any subsequent rigid head motion is taken care of by the intra-sequence registration process described above.

We can now compute the values of the registered points P_r as:

$$P_r(t) = T_E(T_A(P, t)) \quad (5.10)$$

²In fact, any frame that displays a neutral face could be used. We only need to know in advance which frame is neutral. Finding that frame automatically would be a study on its own.

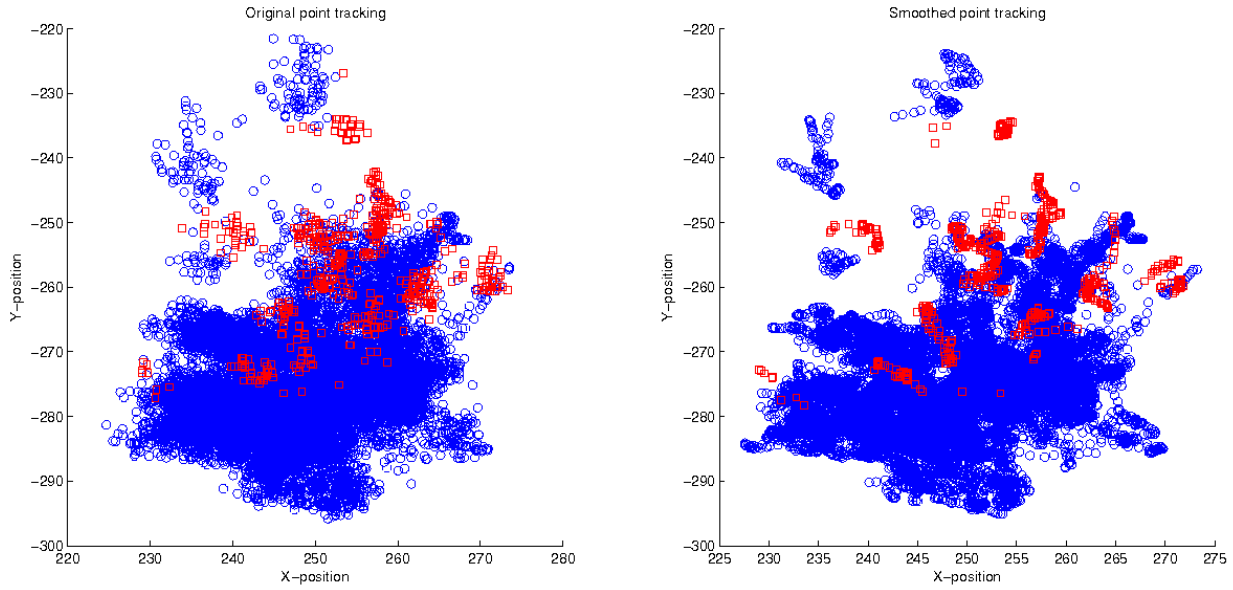


Figure 5.3: Noise reduction by applying a temporal filter to two features relevant for detection of Action Unit 1 (inner brow raise). The x axis represents the x-position of point D (right inner brow) and the y-axis represents its y-position. Circles are examples from frames where the AU was not active, squares are from frames where the AU was at its peak. The left figure shows the unfiltered features, while the right figure clearly shows a reduction in noise and clearer spatio-temporal patterns.

5.2.3 Smoothing the tracking data

The tracked points returned by the PFFL tracker contain some random noise around their true position. This is characteristic for particle filtering. Therefore, we apply a temporal smoothing filter on this tracked and registered data P_r to arrive at a feature set P' that contains less noise:

$$\mathbf{p}'_i(t) = \frac{1}{2w_s + 1} \sum_{t-w_s}^{t+w_s} \mathbf{p}_i^r(t) \quad (5.11)$$

where t denotes the frame number and \mathbf{p}' and \mathbf{p}^r are elements of the collections P' and P_r , respectively. The window sidelobe size w_s to which we apply the temporal smoothing (i.e., the number of frames left and right of the centre of the window) was chosen after visual inspection of the smoothed tracker's output. For the experiments discussed in this thesis, $w_s = 1$ has been chosen. Figure 5.3 clearly shows the noise reduction achieved by applying the temporal filter.

5.3 Mid-level parameter extraction

The registration and smoothing pre-processing steps return for every image sequence a set of points P' , with dimensions $l \times 20 \times 2$, where l is the number of frames of the input image sequence. From this tracking data, we will compute three different feature sets F_S , F_D and F_W .

5.3.1 Single frame based features

The most basic features that can be computed from the tracked point information are the positions of the points and the distances between points. We also compute the angles that the line connecting two points makes with the line $y = 0$ (the horizontal axis). As the faces are already registered before we compute the features, all points in a neutral face should be more or less at the same position for each subject. When a facial point moves, it will be found in some areas typical for that AU. For instance, when someone smiles (AU12) we can expect point I (the right mouth corner) to be found in an area above and to the right of the area where we can expect the point in its neutral state. Thus we can expect to be able to recognise facial actions using these features.

For each point \mathbf{p}_i , where $i = [1 : 20]$, the first two features are simply its x and y position. We compute the features \mathbf{f}_1 and \mathbf{f}_2 for every frame t :

$$f_1(\mathbf{p}_i, t) = p_{i,x,t} \quad (5.12)$$

$$f_2(\mathbf{p}_i, t) = p_{i,y,t} \quad (5.13)$$

For all pairs of points $\mathbf{p}_i, \mathbf{p}_j$, $i \neq j$ we compute in each frame two features:

$$f_3(\mathbf{p}_i, \mathbf{p}_j, t) = \|p_{i,t} - p_{j,t}\| \quad (5.14)$$

$$f_4(\mathbf{p}_i, \mathbf{p}_j, t) = \arctan \left(\frac{p_{i,y,t} - p_{j,y,t}}{p_{i,x,t} - p_{j,x,t}} \right) \quad (5.15)$$

where \arctan is the modified inverse tangent function that corrects for the quadrant a point is in (i.e. solves the arctangent problem). Feature \mathbf{f}_3 describes the distances between two points \mathbf{p}_i and \mathbf{p}_j , and

feature f_4 describes the angle that the line connecting p_i with p_j makes with the horizontal axis. The features f_1 and f_2 are computed for all points $p_i \in P'$, while the features f_3 , and f_4 are computed for all possible combinations of points (190 combinations in our case).

The collection of features $\langle f_1 \dots f_4 \rangle$ constitutes the set of single frame based features F_S with dimensionality $D_{F_S} = 2 \times 20 + 2 \times 190 = 420$. All features in this set are computed using only one frame; the frame t for which the features are calculated. This means that the set F_S is purely static and captures no temporal dynamics whatsoever.

5.3.2 Inter frame based features

The features $\langle f_1 \dots f_4 \rangle$ contain only the information about the positions of the points, the distances between them and the angles they make with the horizontal at the current instance in time. No information about the relation of these measurements to their values in a frame displaying a neutral expression is encoded. Neither do they encode any information about the rate of change of the values of these features in consecutive frames (e.g. the speed of a point). To capture this temporal information, we create a new set of features based on the single frame based features described above.

First, we compute features that describe how much the feature values have changed, relative to their value at the neutral frame, that is, the frame in which no facial expression is shown. In theory this could be any frame, but for now we will assume that this is the first frame of an image sequence. We do so using the *difference function* $\kappa(\mathbf{x}, t)$:

$$\kappa(\mathbf{x}, t) = x_t - x_1 \quad (5.16)$$

where \mathbf{x} is a time sequence and x_t its value at time t . Using this definition we compute the following features:

$$\langle f_5(t) \dots f_8(t) \rangle = \langle \delta(f_1, t) \dots \delta(f_4, t) \rangle \quad (5.17)$$

And to determine the rate of change of the feature values at a given time instance t , we compute their first derivative with respect to time. Because we are working with discrete data, this becomes:

$$\frac{d(\mathbf{x}, t)}{dt} = v(x_t - x_{t-1}) \quad (5.18)$$

where v is the framerate of the corresponding recording and we use this definition to compute the features:

$$\langle f_9(t) \dots f_{12}(t) \rangle = \langle d(f_1, t)/dt \dots d(f_4, t)/dt \rangle \quad (5.19)$$

This results in a feature set $F_D = \langle f_5 \dots f_{12} \rangle$ with dimensionality $D_{F_D} = 2 \times (2 \times 20 + 2 \times 190) = 840$. The feature set F_D compares the change in feature values between the current frame and a single previous frame. This means that the set captures some of the temporal dynamics of the facial point motion.

5.3.3 Local time parameterised features

The time instance based feature set F_S is computed using the tracking information of one frame (features $f_1 \dots f_4$) or at most two frames (features $f_5 \dots f_{12}$). As such, their temporal scope is limited. Not only does this mean that it is hard to model dynamics of facial actions on longer time scales, but it also makes the system very sensitive to inaccuracies of the tracker that escaped the correction by the smoothing process (see section 5.2.3 above). An error in a single time instance will likely result in an outlier in our data representation. This, in turn, will lead to lower classification accuracy.

A second problem with the instance based features is that they fail to capture enough information about the evolution of the facial point motion over a longer period of time. We believe this information can be very useful for AU detection. It can encode whether a point is moving up for a number of consecutive frames, or is remaining at the same location for a period of time. We believe it can be of special importance for the next stage: the detection of the temporal segments of AUs, because it can encode very well the shape of the path of a point when an AU changes from a neutral phase (points remain at the same place) to an onset phase (points are moving away from their neutral position). Therefore we propose to describe the path of the features in a time window of duration T_w frames with a p_w -th order polynomial. We choose T_w and p_w such that with a given framerate we can accurately describe the feature-path shape in the fastest facial AU (AU45, a blink). For our data, recorded at 25 or 30 frames per second, this results in $T_w = 7$ and $p_w = 2$.

Thus, we create a new feature set F_W by computing the mid-level parameters $f_{13} \dots f_{24}$. These are found to be the values that make the following polynomial fit best:

$$f_k = f_{3k+10}t^2 + f_{3k+11}t + f_{3k+12}, k \in [1 \dots 4] \quad (5.20)$$

Thus, the features f_{13} , f_{14} and f_{15} are found for $k = 1$ as the parameters that fit the polynomial $f_1 = f_{13}t^2 + f_{14}t + f_{15}$ best. This results in a feature set $F_W = \langle f_{13} \dots f_{24} \rangle$ with dimensionality $D_{F_W} = 3 \times D_{F_S} = 1260$.

In chapter 6 we will perform an extensive evaluation on the three sets of features. We will look at the importance of the sets, both separately and combined, for the recognition of AU activation and for the recognition of the temporal segments of AUs.

Chapter 6

Action Unit activation detection

In the following two chapters we will present our proposed AU recognition system. The system consists of two major consecutive parts, each described in a separate chapter. The first part is the AU activation detection system, which we will describe in this chapter. This part determines for each frame whether an AU is present as well as whether that AU was active during an entire image sequence (AU event detection). The second part, described in chapter 7, relates to the analysis of AU temporal dynamics and determines exactly when an AU begins (onset), when it reaches a peak(s) (apex), when it starts diminishing again (offset) and when it returns to its neutral phase (neutral). We will also refer to this as onset-apex-offset detection. An improved event detection method, based on the temporal analysis of AUs, is presented in chapter 7 as well. Figure 6.1 gives an outline of the final fully automatic system that combines all the system elements described in chapters 4 to 7.

The AU activation detection and AU temporal analysis parts are independent components. When we combine all system components into a fully automated system, however, the AU activation detection sub-system will only call the AU temporal analysis sub-system for AUs that were determined to be active in that particular image sequence. This reduces the computational complexity of the system. But first some words on the way we measure the performance of our system in this chapter and the next, and the way our evaluation procedures are carried out.

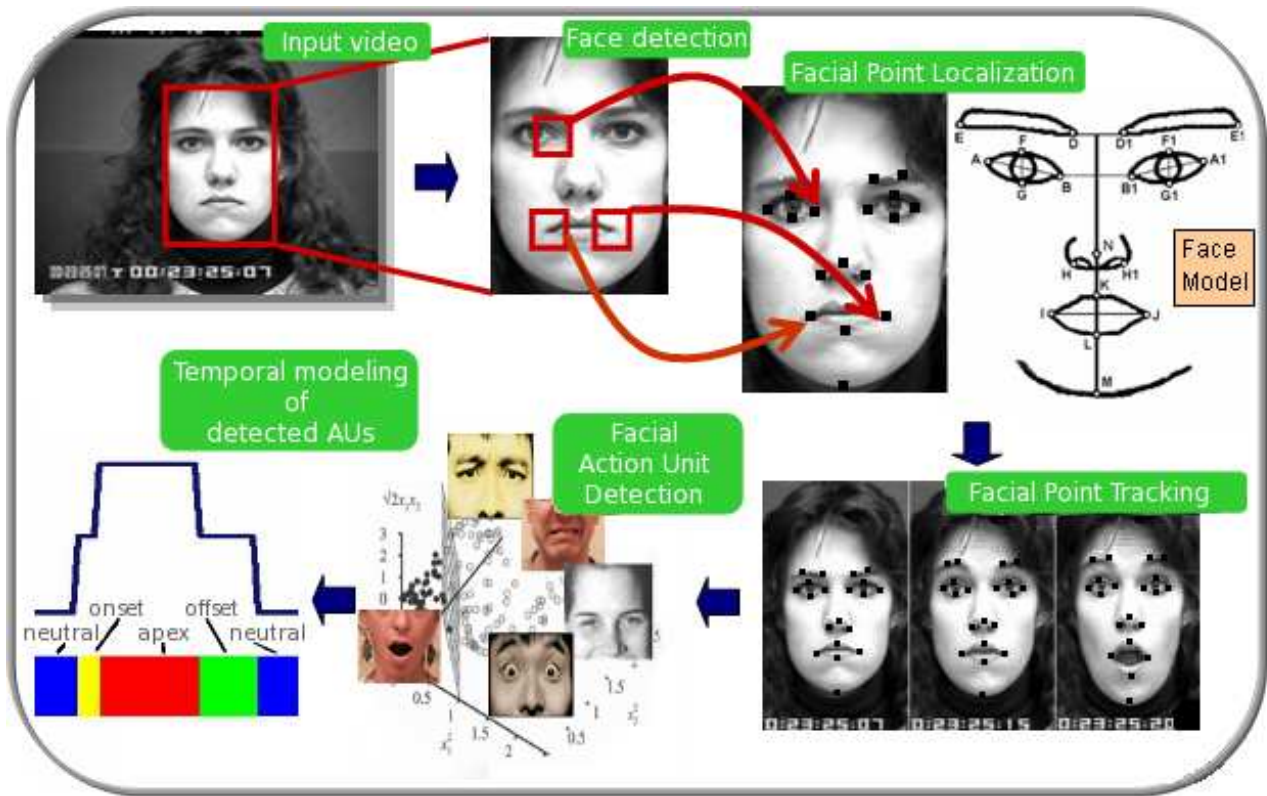


Figure 6.1: Overview of the fully automatic AU analysis system.

6.1 Performance measures

In this work we will carry out a number of evaluation studies. Here we introduce the four performance measures that we will use in these studies. The four measures that we will use are the classification rate (c_r), recall rate (r_r), precision rate (p_r) and the F1-measure (ϕ_1 , a special case of the F-measure that favours recall and precision equally). These measures are defined as follows. Let us define the number of negative examples in the test set as n_n and the number of positive examples in the test set as n_p . Now let us denote the number of examples classified correctly as negative by m_{tn} , (true negatives), the number of examples incorrectly classified as negative by m_{fn} (false negatives), the number of examples correctly classified as positive by m_{tp} (true positives) and the number of examples incorrectly classified as positive by m_{fp} (false positives).

The classification rate c_r measures the overall fraction of correctly classified examples. It is the most intuitive and most (mis)used performance measure. It is defined as:

$$c_r = \frac{m_{tp} + m_{tn}}{n_p + n_n} \quad (6.1)$$

The classification rate is only a useful measure when the test set is balanced: that is, when the sample size of the positive examples is approximately as big as the sample size of negative examples [Sokolova et al., 2006]. Given an unbalanced test set, c_r will give a misleading perception of performance. For example, let us take a test set with $n_n = 990$ and $n_p = 10$. A classifier that would decide that every example presented to it is negative would achieve a classification rate of 0.99, which is nearly perfect, one would be tempted to say. However it is clear that this is not to be considered desired classifier behaviour.

Therefore we introduce the recall and precision rates. The recall rate r_r is a measure of how many elements of one class (usually denoted as the *positive* class) are retrieved by the classifier. In short, it tells us what chance there is that a positive test example will be recognised as such by the classification algorithm. It is used frequently in situations where the data is very unbalanced, as is the case with AU detection. It is defined as the fraction of the true positives over the number of positive examples in the test set:

$$r_r = \frac{m_{tp}}{n_p} \quad (6.2)$$

Still, given an unbalanced test set, we could obtain a classification performance that results in a very high, even perfect, recall, but which is still useless in practice. As a simple intuitive example, take a classifier which predicts all test examples as positive. In this case $m_{tp} = n_p$ and thus $r_r = 1$. However, it is obvious that such a system would not be of any use to us either. Therefore the precision was introduced.

The precision is a measure that tells us how certain we can be that an example that was classified as positive is indeed a positive example. It is defined as:

$$p_r = \frac{m_{tp}}{m_{tp} + m_{fp}} \quad (6.3)$$

Now we have two measures that together give a good indication of the classifier performance. However, it is hard to compare two systems, or the same system trained with slightly different parameters, using two different performance measures. Sometimes the recall will increase, while precision will decrease when we modify a system parameter (or vice versa), and it is therefore hard to tell for which parameters the system performs better.

To solve this problem, the F-measure was introduced. This measure computes a single performance value based on both the system's recall and precision. The general F-measure formulation allows us to set a parameter α_r to indicate whether we prefer recall over precision. It is defined as follows:

$$\phi = \frac{(1 + \alpha_r)p_r r_r}{\alpha p_r + r_r} \quad (6.4)$$

As we do not have a preference for either the precision or recall at this point we use a special case of the F-measure called the F1-measure ϕ_1 . This is a measure of performance that values recall as important as precision, i.e. $\alpha_r = 1$, and is computed as follows:

$$\phi_1 = \frac{2p_r r_r}{p_r + r_r} \quad (6.5)$$

It is this measure that we will use to compare various variants of our proposed system and to optimise for a given parameter. However, to allow a greater insight in the performance of the proposed prototype systems, we will at times provide the other three measures (i.e. classification rate, recall and precision) as well.

6.2 Evaluation procedures

Even though the MMI facial expression database contains the largest amount of single AU activation facial expressions currently available in the world, we still have preciously little data. We will show in section 6.4.1 that adding more data to our training set would increase the accuracy of our system. Because of this, we use n -fold cross validation to evaluate our algorithms. With n -fold cross-validation we partition the dataset in n parts and have n evaluation rounds. In each round, one part is left out of the training set. The classifier is trained on all other parts and tested on the part that was left out. This way, we test on each datum exactly once. Results are reported as the performance measure averaged over the n folds.

As stated above, we have preciously little data. To be able to use as much of this data as possible for training of our systems, we want to choose a large value for n . This results in small individual partition sizes, and consequently in a great amount of data to train the classifier on. The downside of having a large n is that the evaluation takes a long time. We have to train the classifier n times, and

a large n means a large training set which in turn means longer training times.

Another related issue is the way in which the data is partitioned. It has been shown that person-dependent systems, where data from a subject are present in both the training set and the test set, outperform person-independent systems [Cohen et al., 2002]. However, the system we intend to build should be able to recognise expressions of any person, including people who were not encountered previously. To report on person-independent system performance, it is important to partition the data in such a way that there are no data from one subject in both the training and the test set. Keeping in mind the scarcity of our data, we thus partition our dataset into the maximum n possible subsets, where n is the number of subjects in the dataset. Note that the partitions may have varying sizes, as our data contain more examples of some subjects than of others. We call this leave-one-subject-out cross-validation.

As we will see later, the classifiers that we will use have a small number of parameters to set while training the classifier. Of course, we want to find the optimal parameters for our problem in terms of classification performance. It is important, therefore, that the parameter optimisation process is done completely independently of the test set. Otherwise, our results will seem overly optimistic (see [Wessels et al., 2005]). We therefore employ a separate 3-fold cross validation loop each time we train a classifier in which we search for the optimal parameter. The training data is split in three subsets, two of which are used to train a classifier and the third is used to test the classifier. We evaluate the classifier performance for different values of a parameter, using gradient descent to quickly converge to the parameter for which the optimal results are achieved. We repeat this process for each of the three folds, usually resulting in a slightly different value of the parameter for which performance was maximum. The final chosen parameter value is the average over the values found for the three folds. Parameters are assumed independent and there is no specific order in which the various parameters are optimised. In all the reported evaluation studies in this chapter we optimise for all unknown parameters in this way.

Finally, we should say something about the statistical significance of our results. As Mitchell explains [Mitchell, 1997], in order to use the t-test or to compute the standard deviation of our outcomes, we need a sample size of at least 30 independently identically distributed (i.i.d.) examples in each test set. As the frames within one video are highly correlated, this means that we theoretically have as many samples as we have videos. To make things worse, for computing the recall and precision, only the true positive examples count. As we will see, given the size of our dataset we do not even come

close to the required 30 i.i.d. test examples when computing recall or precision. Often we only have 1 or 2 positive examples in the test set. What this means is that we cannot give standard deviations or perform t-tests on the recall and precision rates, but only on the classification rate.

6.3 Action Unit activation detection

Our proposed method for the recognition of AUs within an image sequence is a three-step process. First, we use the GentleBoost algorithm to select the most informative features, thus reducing the problem space which in turn increases the classification accuracy [Bartlett et al., 2004]. Next we use Support Vector Machines (SVMs) to score AUs on a frame-by-frame basis, using the selected features only. Finally, based on this frame-based AU activation detection, we decide whether a particular AU is shown in the input image sequence or not. This last step is called AU event-detection, and it is done using an adaptive threshold that counts the number of active frames in an image sequence.

6.3.1 Feature Selection

Although SVM classifiers are able to learn a classification function very well even with very little training data, classification performance decreases when the dimensionality of the training set is too large. To be more precise, if we have a training set D with n_f features and n_s instances, then if $n_f > n_s$, it is possible to uniquely describe every example in the training set by a specific set of feature values [Vapnik, 1995]. Our training set typically consists of some 250 examples (image sequences) of which as few as 6 are from the class we wish to detect (*positive* examples) while all other examples are from non-target classes (*negative* examples). Considering the fact that the dimensionalities of our feature sets, D_{F_S} , D_{F_D} and D_{F_W} , are 420, 840 and 1260 (see section 5.3), respectively, we are indeed in danger of over-fitting to the training set. One way to overcome this problem is to reduce the number of features used to train the SVM using feature selection algorithms.

Boosting algorithms such as GentleBoost or AdaBoost are not only fast classifiers, they are also excellent feature selection techniques. Bartlett et al. [Bartlett et al., 2004] reported that SVMs trained on a subset of features selected by a boosting algorithm perform significantly better. They also showed that not all feature selection techniques are suitable for use with SVMs. For instance, the use of PCA as a feature selector for SVMs was even shown to have a detrimental effect to the final SVM classification

performance.

In our study we use the GentleBoost algorithm as a feature selector preceding a SVM classifier. The performance of GentleBoost is reportedly better than that of AdaBoost, both in convergence time and classification accuracy [Torralba et al., 2004, Lienhart et al., 2003]. An early-stage empirical study confirmed this for our problem. We evaluated the performance increase on 153 examples taken from the Cohn-Kanade Database. There was an 18.9% increase in recall when we used Gentle-SVMs compared to conventional SVMs.

In feature selection by GentleBoost, at each stage a weak classifier is trained on a subset of the data consisting of a single feature. So, if we combine the three feature sets described in section 5, we have 2520 features, and thus 2520 possible weak classifiers. The nature of boosting techniques such as GentleBoost or AdaBoost ensures that new features are selected contingent on the features that have already been selected, eliminating redundant information with each training round.

The feature selection procedure based on a boosting algorithm goes as follows: The algorithm picks the weak classifier (based on a single feature) that separates the examples of different classes best given the current sample weights w , and then boosts the weights to weigh the errors more (so-called importance resampling). The next feature is selected as the one that gives the best weighted performance on the data set given the updated weights. It can be shown that at each step the chosen feature is highly uncorrelated with the previously selected features [Friedman et al., 2000]. The difference between GentleBoost and AdaBoost lies in the reweighting scheme, which is a linear function of the error for GentleBoost and a half-log function of the error for AdaBoost. Therefore AdaBoost can over-emphasise the errors. This can lead to lower performance rates, especially on data for which the Bayes error is significant [Freund and Schapire, 2000]. A formal description of the GentleBoost algorithm has been given in section 4.4 of this thesis.

The performance measure used by GentleBoost to learn the weak classifiers and evaluate their performance is the classification rate (i.e. fraction of correctly classified examples, regardless of their class). As stated above, our dataset is extremely unbalanced, with approximately 15 times as many negative examples as positive examples. Therefore, care must be taken to avoid that GentleBoost favours correct classification of the negative examples, as this would lead to a system that would be practically unable to detect positive examples¹. The solution lies in the way the boosting algorithm evaluates its performance in every round eq.(4.10).

¹In other words, the system would have a very low recall.

Let us denote with I_p the indices of the positive examples and with I_n the indices of the negative examples in our data set. Let the weights of positive examples be denoted as $w_p = w(I_p)$ and the weights of negative examples as $w_n = w(I_n)$. Now we initialise w in such a way that $\sum w_p = 0.5$ and $\sum w_n = 0.5$ and thus $\sum w = 1$. We call this *equal-class weighting*, as opposed to giving each training example the same weight, which we call *equal-instance weighting*. Moreover, we insist that in the first round of boosting, $(\forall i, j \in I_p) w_i = w_j$ and $(\forall i, j \in I_n) w_i = w_j$. Using this initialisation of the weights, we ensure that GentleBoost endeavours to have an equal error on the positive examples and on the negative examples.

To evaluate the effect of initialising GentleBoost with equal-class weighting, we performed two AU-activation detection tests: In the first test we initialise GentleBoost using equal weights for all examples (standard initialisation with *equal instance weighting*). In the second test we initialise GentleBoost using equal-class weighting. To reduce the amount of time needed to evaluate the effect of using equal-class or equal-instance weighting, the performance was measured on a subset of 10 AUs: AU1, AU2, AU4, AU5, AU9, AU12, AU15, AU18, AU20 and AU22. This set of AUs was chosen to reflect a wide range of AUs. It consists of both upper-face, lower-face and ‘miscellaneous’ AUs and some of the AUs are very easy to recognise (e.g. AU2) and others very difficult (e.g. AU15). We will use this set for more tests and will refer to it as the *evaluation data set*. For each AU, we used half of the data to train GentleBoost and the other half to test it. Figure 6.2 shows the results of this test, which are explained in greater detail below.

As we can see from Fig. 6.2, it is very important to use the appropriate performance measure. If we would only regard the classification rate, we would incorrectly conclude that equal-instance weighting provides optimal results. The figure also shows that the performance measured on the training data is misleading. The bottom-right figure finally shows us that the F1-measure on the test data is actually greatest when we use equal-class weighting.

Figure 6.2 also points out another long-debated [Freund and Schapire, 1999, Jiang, 2004] issue of training boosting classifiers. As the lower-right graph clearly shows, on our data GentleBoost is prone to overtraining. The figure shows us that, to obtain the best results, the algorithm should have stopped training after 8 rounds. Unfortunately, there is no way for us to know *a priori* after how many rounds training should be stopped. We investigated the issue further by forcing the boosting algorithm to finish after a fixed number of rounds (early stopping), and by subsequently using only those selected features in the full frame-by-frame AU-activation detection system described in the following sections.

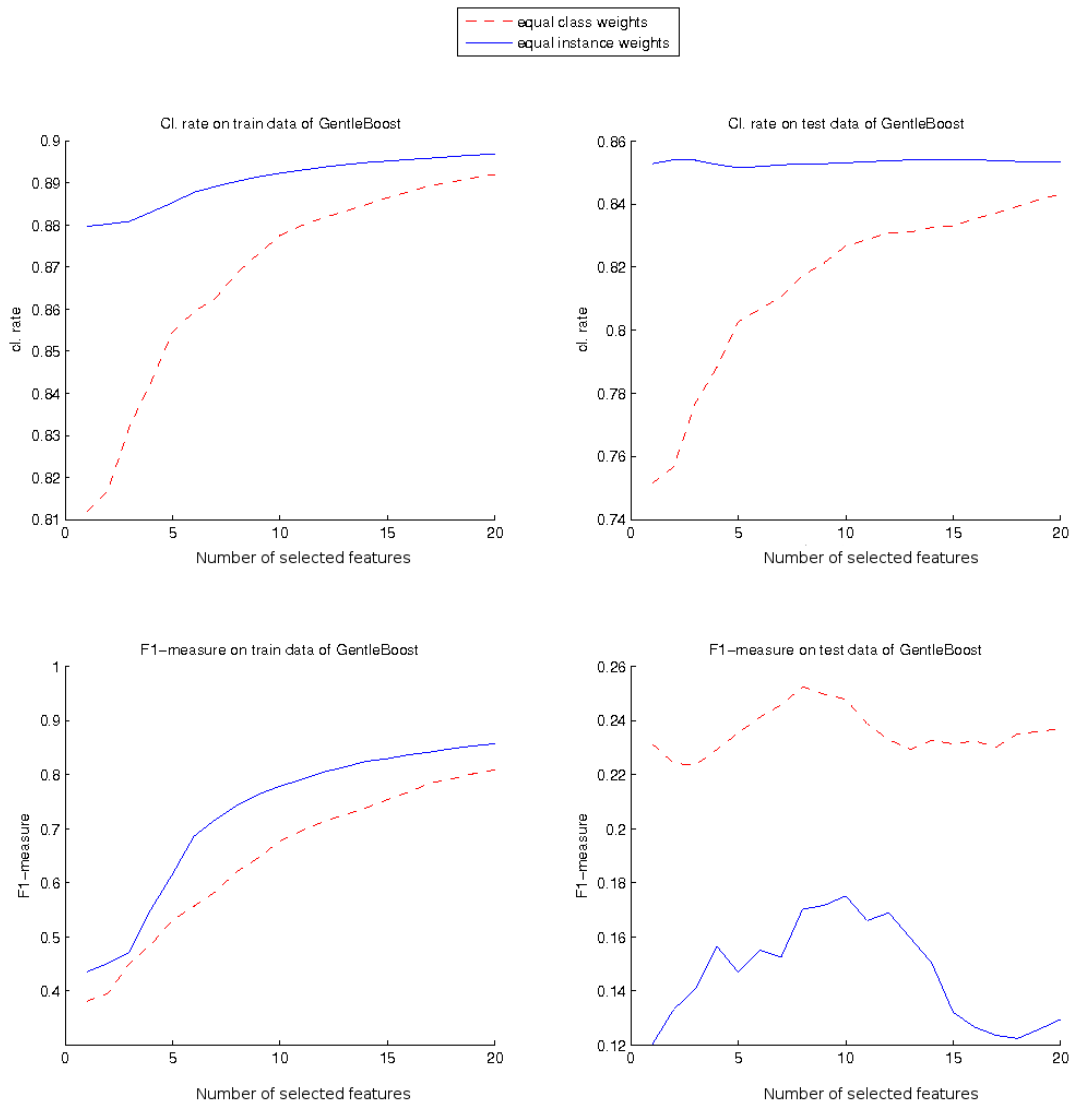


Figure 6.2: Performance of GentleBoost with different initialisations of the weights, measured in classification rate and F1-measure.

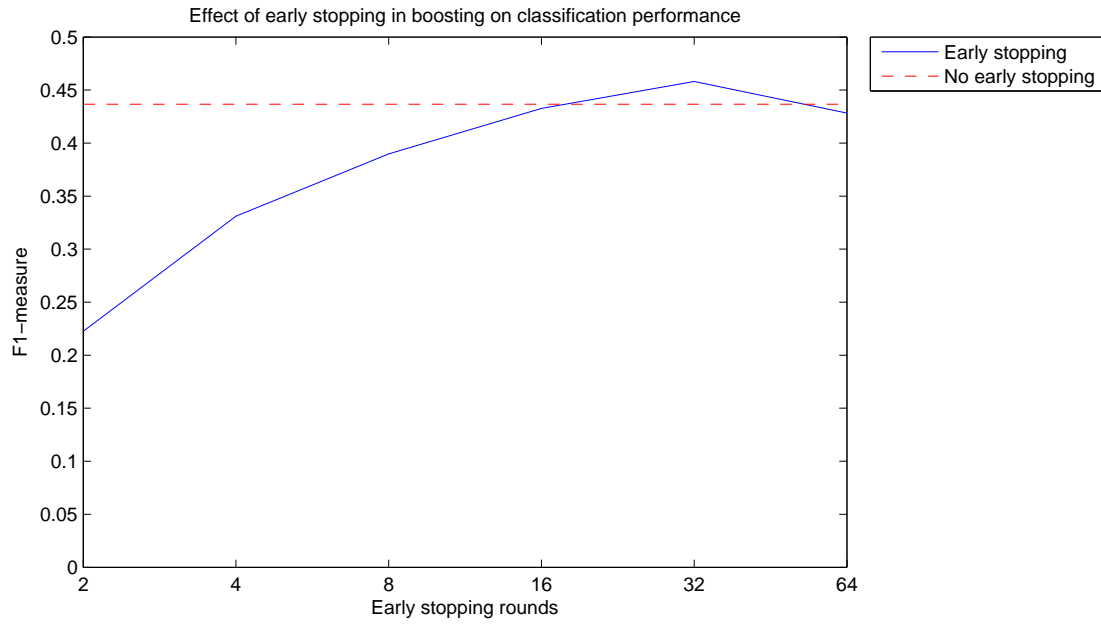


Figure 6.3: The performance of AU-activation recognition for different fixed numbers of features selected by GentleBoost.

The results of this 10-fold cross validation test on the *evaluation data set* are shown in figure 6.3. The solid line shows the AU-activation detection performance for different early-stopping values while the dotted line shows the performance for AU-activation detection using feature selection without early stopping. The figure seems to indicate that some overfitting does occur, which is in agreement with the findings reported in [Jiang, 2004]. The figure shows that an optimal number of boosting rounds exists and thus that early stopping could be beneficial to classifier performance. Yet the difference between this optimal performance and the performance without early stopping is smaller than one might expect from inspection of figure 6.2.

This optimal number of rounds depends on many factors, such as the definition of the features, the amount of training data, and the amount of noise introduced by the (sub)systems. Finding this optimal value would require a separate cross validation loop each time a classifier is trained. Therefore, taking into account that the performance increase is very small, we decided not to perform early stopping in our systems.

6.3.2 Support Vector Machine Classification

Support Vector Machines (SVMs) have proved to be very well suited for classification tasks such as facial expression recognition because, in general, the high dimensionality of the input feature space does not affect the training time, which depends only on the number of training examples [Vapnik, 1995]. They are non-linear, generalise very well and have a well-founded mathematical basis. The essence of SVMs can be summarised as follows: maximising the hyperplane margin, mapping the input space to a (hopefully) linearly separable feature space and applying the ‘kernel trick’ to efficiently solve the first two steps, by implicitly mapping the samples to hyperspace when computing the distance between two examples.

Maximising the margin of the separating hyperplane \mathbf{h} results in a high generalisation ability. To achieve this, one has to find the examples that span the hyperplane (the support vectors, SVs) that maximises the distance between the SVs and \mathbf{h} . This involves finding the nonzero solutions of the instance weights γ_i of the Lagrangian dual problem [Vapnik, 1995]. This dual problem is a quadratic programming problem which can be solved efficiently. The instance weights γ_i are mostly zero. The examples for which γ_i is not zero are exactly the SVs and this sparse set of vectors is the only part of the training data that has to be stored to define the trained classifier. Generally speaking, a very sparse set of SVs means high generalisation power.

Having found the support vector weights γ_i and given a labelled training set $\{\mathbf{x}, \mathbf{y}\}$ of features \mathbf{x} and class labels \mathbf{y} , the decision function in input space is:

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^l \gamma_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b \right) \quad (6.6)$$

where b is the bias of the hyperplane, $\langle \mathbf{x}, \mathbf{x}_i \rangle$ is the inner product of the test example \mathbf{x} and the i th training example \mathbf{x}_i with corresponding class label y_i , and l is the training sample size. The function $\text{sgn}(y)$ returns the sign of y , i.e. either -1 or 1.

Equation (6.6) is a linear decision function. Of course, most real-world problems are not linearly separable in input space. To overcome this problem, we map each input sample \mathbf{x}_i to its representation in the *feature space*: $\Phi(\mathbf{x}_i)$. In this space we hope that the decision function *is* linear, so we can apply our algorithm for finding the maximal margin hyperplane in it. This requires the computation of the dot product $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle$ in *feature space*, which is a high-dimensional space. These expensive

calculations are reduced significantly by using a Mercer kernel K . These kernels have the property

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle = K(\mathbf{x}, \mathbf{x}_i) \quad (6.7)$$

To find the non-linear decision function for $f(x)$ that defines \mathbf{h} , we apply our decision function (6.6) directly on $\Phi(\mathbf{x})$. Substituting the inner product in equation (6.6) with the Mercer kernel (6.7), the decision function directly becomes:

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m y_i \gamma_i K(\mathbf{x}, \mathbf{x}_i) + b \right) \quad (6.8)$$

In many of the studies presented in this thesis, we will use a combination of GentleBoost feature selection and SVM classification. For this combination the term gentleSvm was coined by [Bartlett et al., 2004].

6.3.3 Cascade Support Vector Machines

Although from an AU event detection point of view we only have around 250 examples in our training set (i.e. 250 image sequences), the gentleSvm system we propose recognises AUs in each frame separately. As such, it uses each frame as a training instance, instead of an entire image sequence. With a typical image sequence length of a hundred frames, this means we have in the order of $n_s = 2.5 \times 10^4$ training examples to deal with.

Theoretically, SVM training time scales as n_s^3 . However, most SVM implementations use Sequential Minimal Optimisation [Cristianini and Shawe-Taylor, 2000], a quadratic programming solution algorithm, which reduces the training time required to approximately n_s^2 . Nonetheless, that amount of training time is still problematic, especially if we keep expanding our training set. Another problem that we encountered when training our system relates to memory requirements. Although a trained SVM only retains training examples for which $\gamma_i \neq 0$, it still needs to load all the training examples in order to train the system. Both problems need to be addressed to be able to evaluate our proposed AU recognition system and to train our final feed-forward AU detector. After testing various data reduction techniques², we decided to use Cascade SVMs [Graf et al., 2004].

²Unfortunately, due to a lack of space we will not divulge here on the matter

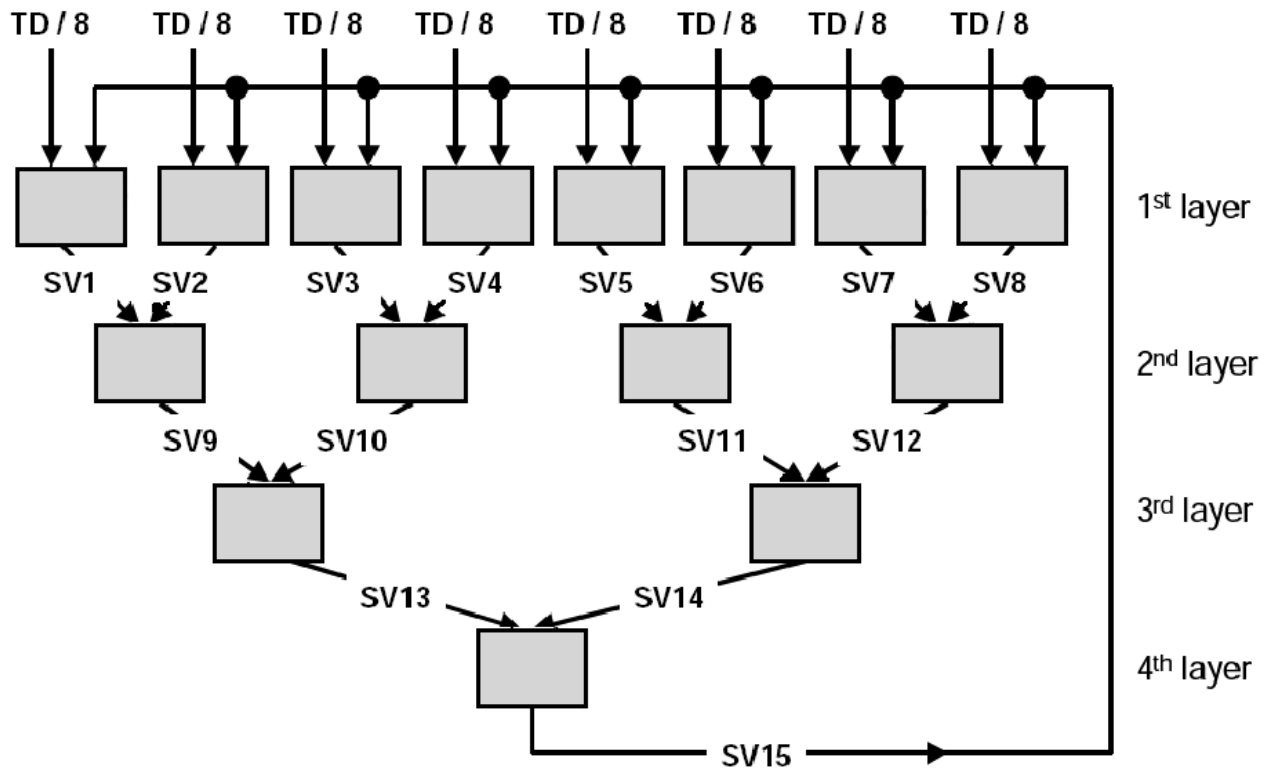


Figure 6.4: Schematic overview of the Cascade SVM

Cascade-SVMs were proposed by a group led by Vapnik, the creator of the original SVM. The system is schematically shown in Fig. 6.4. A cascade SVM with n_c cascades has $n_c + 1$ layers and splits the training set that serves as input to the SVMs of the first layer in 2^{n_c} parts. When the cascade-SVM is trained for the first time, the first layer is trained using the split training set only. In the second layer and beyond, each SVM is trained using the Support Vectors found by its two parent SVMs. This continues until we reach the final layer.

The number of cascades n_c to use is the most important model parameter. Figure 6.5 shows us that we have to make a trade-off between the classifier accuracy on the one hand and the training time/memory requirements on the other. Considering the specifications of our lab computers and the size of our dataset, we set $n_c = 2$. This means that our training data is split in 4 equally sized parts. In this setup the training data just fits in the machine memory, as we only need a quarter of the original memory requirements. In a worst case scenario, that is, if each SVM in every layer would select all training examples to be used as support vectors, training time would actually increase and we would have no relaxed memory requirements (the final layer would ultimately train on the original training set). This is of course not a realistic scenario and in practice we observed an almost 4-fold

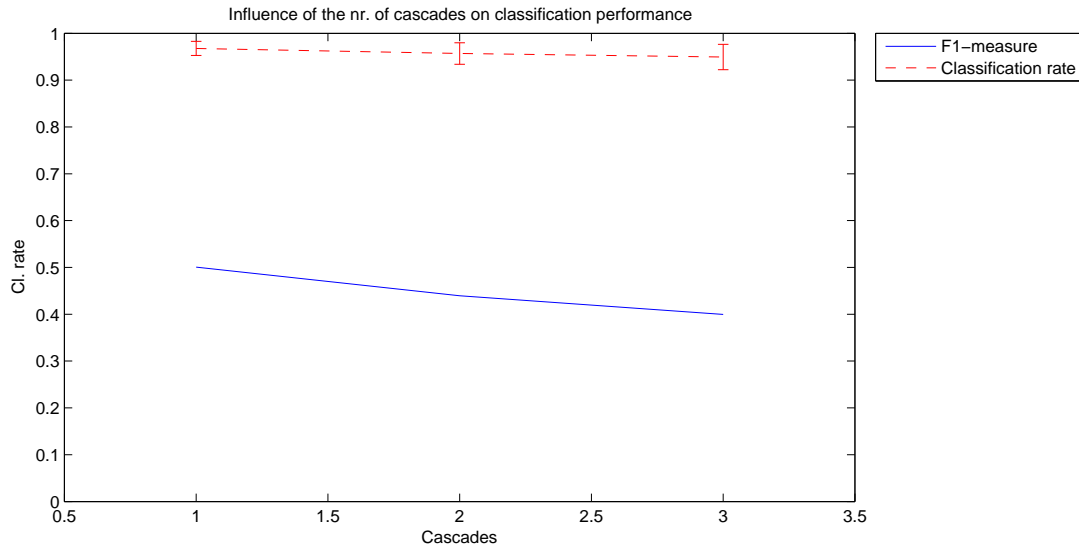


Figure 6.5: Effect of the number of cascade loops on the classification performance.

speeding up of the training algorithm, and memory usage never exceeded that required to train the first layer.

If we would traverse this cascade only once, convergence with the optimal SVM (optimal in terms of classification accuracy; the case where we do not use Cascade-SVM and train only one SVM using all data) is not guaranteed. Intuitively we can see this because some training data passed in to the third SVM of the input layer might not be found to be Support Vectors, while they could have been selected if they were presented to an SVM together with the data of the first SVM. Therefore we use a feedback loop, where we add the Support Vectors selected by the final layer to each of the training data parts that were used to train the first layer and we traverse the cascade again. Graf et al. [Graf et al., 2004] have proven that in this way, the cascade-SVM will eventually converge with the optimal SVM. However, it remains unclear how many feedback rounds would be necessary. Also, [Graf et al., 2004] showed that often the feedback rounds were not necessary at all and that the cascade-SVM was trained optimally in the first round. To see how many feedback loops we needed, we performed a test on the *evaluation data set*. The results are shown in figure 6.6. As we can see, the number of feedback loops used does not significantly impact the classification results. We decided to use two feedback loops, because the mean F1-measure was highest for this value.

Each SVM in the first layer of a cascade-SVM is trained with a maximum of $n_s/2^{n_c}$ samples and it is reasonable to expect that any SVM will select fewer than half of its input data as SVs. As only one SVM needs to be trained at any time, the memory requirements advantage is evident. The training

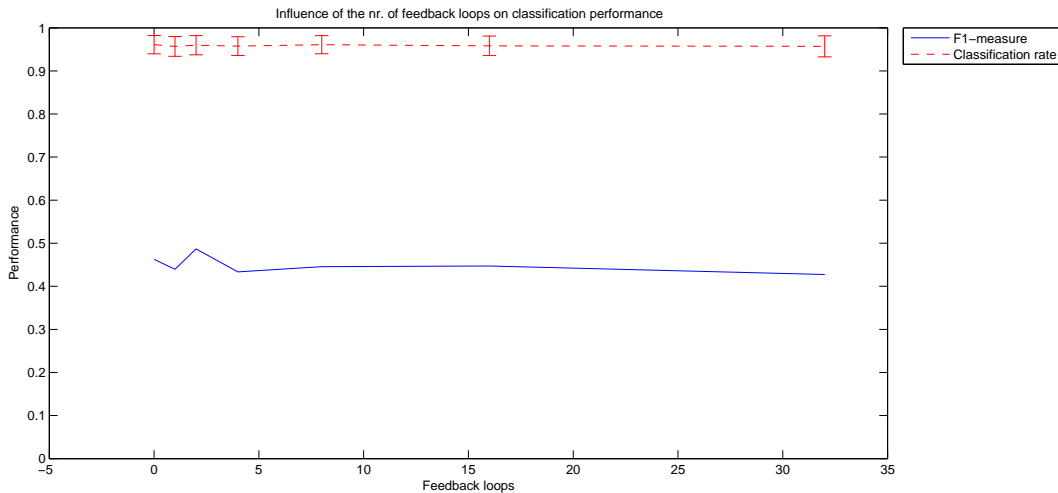


Figure 6.6: Effect of the number of feedback loops on the classification performance.

time advantage is less evident, as it depends on the number of Support Vectors that survive after each layer and the number of iterations for which the cascade-SVM is trained. However, as stated above we observed about a fourfold reduction in training time with two cascades. Also, the cascade-SVM architecture allows for distributed computing, as SVMs could be trained in parallel on different machines.

6.4 Evaluation of AU activation detection

6.4.1 Frame-based AU detection results

We performed four studies to evaluate the performance of our frame-based AU activation detection approach. In the first study, we evaluate the information contained in the various feature sets, F_S , F_D and F_W , defined in section 5.3. Based on this study we decided which features to use for all other tests. The other two studies are a complete leave-one-subject-out cross-validation studies on 244 sequences from the MMI-Facial Expression Database and 153 sequences from the Cohn-Kanade database, respectively. These studies provide a good indication of how well the system can perform when it is trained for a specific application. It is also a reasonable way to compare the performance of our system to that of other systems presented in the literature. The fourth study investigates how well our method is able to generalise with respect to completely novel data, recorded under very different conditions. We train our system using the data from the MMI-Facial Expression Database (the set

Feature sets	Cl. Rate	F1-measure
F_S	0.952	0.381
F_D	0.958	0.447
F_W	0.951	0.395
$\{F_S, F_D\}$	0.959	0.449
$\{F_S, F_W\}$	0.951	0.396
$\{F_D, F_W\}$	0.953	0.438
$\{F_S, F_D, F_W\}$	0.956	0.467

Table 6.1: Performance of frame-based AU activation detection for various feature sets on MMI database.

used in the first study) and test it on data from the Cohn-Kanade Database (the set used in the second study), and vice versa.

Feature set evaluation

In section 5.3 we introduced three different sets of features: the purely static feature set F_S , the feature set based on differences between two frames F_D and the local time parameterised feature set F_W . We want to know the relative importance of these feature sets to the problem of AU activation detection. Therefore we have performed seven 10-fold cross validation studies on the *evaluation data set*, one for each possible combination of feature sets. We used a cascade-SVM with 2 cascades and used GentleBoost as a feature selector. Performance was measured per frame, i.e. we counted correctly classified frames. The results are shown in table 6.1.

As pointed out earlier, the classification rate does not provide much insight. The F1-measure ϕ_1 however, does indicate some important points. As was expected, the purely static feature set F_S performs worst. The set of local time parameterised features F_W performs only marginally better. When combining F_W with either F_S or F_D , no improvement is noticed either. But when combined with both F_S and F_D the best results are obtained. However, the power of F_W lies in describing local time-variations in the positions and velocities of points or in the distances between points. This is probably of greater importance to the recognition of the temporal phases of AUs, where the changes between phases are signified by these local time-variations, than to the activation detection of AUs. In section 7 we will measure the performance on temporal segment detection and we expect to see a greater influence of the local time parameterised feature set F_W .

The best performing feature set is the double-frame based feature set F_D , which captures both the speed of points and the deviation in location and distances between points relative to a neutral frame.

It achieves a ϕ_1 of 44.7%, compared with a score of 46.7% for all feature sets combined.

To get a deeper understanding of the information contained in each of the feature sets, we also looked into which features were selected for AU activation detection. Table 6.2 lists for all AUs that we can detect in the MMI database the four most important features for the problem of AU activation detection. To make the table entries better readable for humans, we decided not to use the feature names $\{f_5 \dots f_8\}$, but instead use their functional representations $\delta(f_i, t)$ and $df_i(t)/dt$ (see section 5.3).

As we can see, the selected features make a lot of sense. For instance, the first feature selected for activation detection of AU1 – the inner eyebrow raiser – is the distance between the left inner eye corner and the left inner eyebrow corner. The inner eye corners are stable points, that is, their position never changes due to facial expressions. As such, they are a very good reference point to measure distances against. The distance between the inner eye corners and the inner eyebrow corners is obviously an indication for a brow raiser. For AU12, a smile, the most important feature was found to be the change in the angle defined by the line connecting the points on the nose and the left mouth corner and the horizontal line. This again makes a lot of sense, as we expect that when we smile, the mouth corner moves away from its neutral position in a circular path, with more or less the nose as the centre of that circle.

As expected, there are very few features selected that encode velocity of acceleration, i.e. time derivatives of features, or the higher order polynomial parameter features. Only 4.7% of the selected features encoded velocity and 6% acceleration. While not increasing recognition accuracy significantly unless combined with both F_S and F_D , the feature set F_W still accounts for 21.4% of the selected features. This, plus the increase in performance when added to F_S and F_D , leads us to conclude that the Local Time Parameterisation features (F_W features) do have added value to recognise facial expressions. Also, we see from table 6.2 that the single-frame based features f_1 and f_2 are hardly ever chosen. To be more exact, of all features chosen, only 3.5% were F_S features denoting either the x - or y -position of a point. Therefore we conclude that it is very hard to recognise facial expressions based on the positions of facial points only.

For the remainder of this thesis, we will use the combination of all feature sets, $\{F_S, F_D, F_W\}$, as this achieves the highest possible performance. However, a word of caution is needed here. Computation of F_W is time-consuming, and the increase in performance that it offers might not always warrant the cost. If a fast, perhaps even real-time, facial expression recognition system is desired then the

AU	Selected Features			
1	$f_3(\mathbf{B1}, \mathbf{D1}, t)$	$\delta(f_3(\mathbf{D1}, \mathbf{N}, t), t)$	$\delta(f_3(\mathbf{A}, \mathbf{M}, t), t)$	$\delta(f_3(\mathbf{D}, \mathbf{N}, t), t)$
2	$f_3(\mathbf{E1}, \mathbf{G1}, t)$	$\delta(f_3(\mathbf{D}, \mathbf{D1}, t), t)$	$\delta(f_4(\mathbf{A1}, \mathbf{B1}, t), t)$	$\delta(f_4(\mathbf{G}, \mathbf{N}, t), t)$
4	$\delta(f_3(\mathbf{E1}, \mathbf{G1}, t), t)$	$\delta(f_3(\mathbf{D}, \mathbf{D1}, t), t)$	$\delta(f_3(\mathbf{E}, \mathbf{G}, t), t)$	$f_3(\mathbf{H1}, \mathbf{K}, t)$
5	$\delta(f_3(\mathbf{F}, \mathbf{G}, t), t)$	$\delta(f_4(\mathbf{A}, \mathbf{G1}, t), t)$	$f_3(\mathbf{H}, \mathbf{N}, t)$	$f_3(\mathbf{E}, \mathbf{F1}, t)$
6	$\delta(f_4(\mathbf{J}, \mathbf{N}, t), t)$	$f_2(\mathbf{G}, t)$	$f_3(\mathbf{E}, \mathbf{H}, t)$	$\delta(f_3(\mathbf{J}, \mathbf{M}, t), t)$
7	$\delta(f_4(\mathbf{E}, \mathbf{G1}, t), t)$	$\delta(f_3(\mathbf{E1}, \mathbf{G}, t), t)$	$\delta(f_3(\mathbf{B}, \mathbf{N}, t), t)$	$\delta(f_4(\mathbf{B}, \mathbf{D}, t), t)$
9	$\delta(f_4(\mathbf{D1}, \mathbf{G}, t), t)$	$\delta(f_3(\mathbf{G}, \mathbf{M}, t), t)$	$f_3(\mathbf{D}, \mathbf{N}, t)$	$\delta(f_3(\mathbf{I}, \mathbf{L}, t), t)$
10	$\delta(f_3(\mathbf{K}, \mathbf{M}, t), t)$	$f_4(\mathbf{J}, \mathbf{M}, t)$	$\delta(f_4(\mathbf{A1}, \mathbf{J}, t), t)$	$f_4(\mathbf{G1}, \mathbf{H1}, t)$
12	$\delta(f_4(\mathbf{J}, \mathbf{N}, t), t)$	$f_{21}(\mathbf{A}, \mathbf{I}, t)$	$f_3(\mathbf{A1}, \mathbf{D1}, t)$	$f_4(\mathbf{A}, \mathbf{I}, t)$
13	$f_4(\mathbf{H}, \mathbf{J}, t)$	$\delta(f_3(\mathbf{L}, \mathbf{M}, t), t)$	$f_3(\mathbf{D}, \mathbf{G}, t)$	$f_{18}(\mathbf{L}, t)$
15	$f_4(\mathbf{J}, \mathbf{L}, t)$	$\delta(f_4(\mathbf{J}, \mathbf{M}, t), t)$	$\delta(f_4(\mathbf{I}, \mathbf{L}, t), t)$	$f_3(\mathbf{B1}, \mathbf{I}, t)$
16	$\delta(f_3(\mathbf{L}, \mathbf{N}, t), t)$	$f_4(\mathbf{B1}, \mathbf{K}, t)$	$f_4(\mathbf{D}, \mathbf{D1}, t)$	$\delta(f_3(\mathbf{L}, \mathbf{M}, t), t)$
18	$f_3(\mathbf{I}, \mathbf{J}, t)$	$\delta(f_3(\mathbf{I}, \mathbf{K}, t), t)$	$f_3(\mathbf{D}, \mathbf{L}, t)$	$f_{24}(\mathbf{G}, \mathbf{I}, t)$
20	$\delta(f_3(\mathbf{I}, \mathbf{J}, t), t)$	$f_3(\mathbf{I}, \mathbf{K}, t)$	$\delta(f_3(\mathbf{B1}, \mathbf{E1}, t)$	$\delta(f_3(\mathbf{B1}, \mathbf{E}, t), t)$
22	$\delta(f_4(\mathbf{K}, \mathbf{J}, t), t)$	$f_3(\mathbf{B}, \mathbf{E1}, t)$	$f_4(\mathbf{D1}, \mathbf{G}, t)$	$f_{21}(\mathbf{J}, \mathbf{M}, t)$
24	$\delta(f_3(\mathbf{I}, \mathbf{N}, t), t)$	$f_3(\mathbf{A}, \mathbf{A1}, t)$	$f_4(\mathbf{E1}, \mathbf{J}, t)$	$\delta(f_3(\mathbf{B1}, \mathbf{D1}, t), t)$
25	$\delta(f_3(\mathbf{K}, \mathbf{L}, t), t)$	$f_2(\mathbf{L}, t)$	$f_{19}(\mathbf{K}, \mathbf{L}, t)$	$f_4(\mathbf{F1}, \mathbf{G}, t)$
26	$\delta(f_3(\mathbf{K}, \mathbf{L}, t), t)$	$f_3(\mathbf{K}, \mathbf{N}, t)$	$f_3(\mathbf{G}, \mathbf{N}, t)$	$f_3(\mathbf{E}, \mathbf{H1}, t)$
27	$\delta(f_3(\mathbf{L}, \mathbf{N}, t), t)$	$\delta(f_4(\mathbf{A1}, \mathbf{L}, t), t)$	$f_{16}(\mathbf{L}, t)$	$\delta(f_3(\mathbf{F}, \mathbf{L}, t), t)$
30	$\delta(f_3(\mathbf{B1}, \mathbf{J}, t), t)$	$f_4(\mathbf{K}, \mathbf{L}, t)$	$\delta(f_1(\mathbf{L}, t), t)$	$f_3(\mathbf{F}, \mathbf{H}, t)$
43	$f_4(\mathbf{F1}, \mathbf{G}, t)$	$f_3(\mathbf{F}, \mathbf{H1}, t)$	$f_{21}(\mathbf{D1}, \mathbf{F1}, t)$	$f_{21}(\mathbf{E}, \mathbf{F}, t)$
45	$\delta(f_4(\mathbf{D1}, \mathbf{F}, t), t)$	$\delta(f_3(\mathbf{F}, \mathbf{G}, t), t)$	$f_{23}(\mathbf{F}, \mathbf{D1}, t)$	$f_{20}(\mathbf{E1}, \mathbf{F1}, t)$
46	$f_3(\mathbf{D1}, \mathbf{G}, t)$	$f_3(\mathbf{A1}, \mathbf{M}, t)$	$\delta(f_4(\mathbf{F1}, \mathbf{H}, t), t)$	$f_3(\mathbf{A}, \mathbf{K}, t)$

Table 6.2: First four features selected by GentleBoost for the activation detection of AUs in the MMI database.

combination of F_S and F_D only is to be preferred.

Frame-based evaluation on MMI database

We tested our system on the MMI-facial expression database for all AUs that we can detect using a geometric feature based approach. The set was created so that it includes for every AU at least 15 examples. For AU13 (a smile with the mouth corners sharply pulled upwards) we could find only 14 examples and for AU46 (wink) only 6. This is because these AUs are very difficult to produce on command.

Some AUs always occur in combination with others. For instance, AU22 which puffs the lips as in pronouncing the word 'flirt', will always cause the lips to part and thus to display AU25. Thus, for some AUs we have more occurrences than for others. The number of videos in which each AU occurs is listed in the second column of table 6.3. Because at this point we are detecting AUs per frame, we have listed the total number of frames in which an AU is active in the third column of table 6.3.

Results are shown in terms of the classification rate c_r , the recall r_r , the precision p_r and the F1-measure ϕ_1 . From the table we can see that there is a large number of AUs that are recognised with quite a high ϕ_1 . AU1, AU2, AU4, AU6, AU13, AU20, AU22, AU27, AU45 and AU46 all have a $\phi_1 > 0.6$, which means that we can expect both a relatively high chance that an active AU is found by the system, and that we can have some confidence that the AUs returned by the system were actually present in the input video.

Some of the AUs that did not score very high were AU10, AU15, AU16 and AU26. For AU10 and AU16, we think the problem lies mainly in the small number of points that define the shape of the mouth. With only one point for each mouth corner, and one point for the upper and one for the lower lip, it is impossible to tell whether the upper lip is a straight line or instead curved. The same holds for the lower lip. As AU10 is caused by muscles that attach halfway between the mouth corner and the midpoint on the upper lip, and AU16 is caused by muscles that attach on the same vertical position on the lower lip, it is exactly this information about the curvature of the lips that is needed to detect AU10 and AU16.

AU15 – lips depressed – is a different case. Although the mouth corners do indeed move down and out a little bit, the most salient change caused by AU15 activation is not in the displacement of the mouth corners but in the appearance change of the mouth corners. A distinct edge, running from the

AU	Examples	Frames	Cl. Rate	Recall	Precision	F1-measure
1	22	1006	0.972	0.679	0.728	0.703
2	25	1092	0.961	0.628	0.629	0.628
4	38	1839	0.942	0.582	0.707	0.639
5	19	874	0.949	0.317	0.375	0.344
6	27	1241	0.952	0.695	0.583	0.634
7	15	772	0.963	0.319	0.510	0.392
9	15	636	0.968	0.503	0.477	0.490
10	17	719	0.955	0.266	0.321	0.291
12	17	1004	0.950	0.548	0.482	0.513
13	14	782	0.974	0.668	0.650	0.659
15	15	854	0.944	0.412	0.344	0.375
16	18	717	0.947	0.230	0.229	0.229
18	16	568	0.974	0.593	0.523	0.556
20	15	871	0.964	0.696	0.554	0.617
22	15	696	0.964	0.536	0.467	0.499
24	15	536	0.955	0.497	0.503	0.500
25	105	5401	0.909	0.810	0.831	0.821
26	32	1597	0.875	0.198	0.179	0.188
27	15	800	0.983	0.720	0.819	0.766
30	15	736	0.972	0.438	0.588	0.502
43	15	750	0.973	0.520	0.657	0.580
45	107	1243	0.956	0.668	0.625	0.645
46	6	130	0.913	0.723	0.667	0.694
Avg:			0.953	0.532	0.541	0.533

Table 6.3: Subject independent cross validation results for AU activation detection per frame on 244 examples from the MMI-Facial Expression Database

mouth corner outwards and downwards is created when AU15 occurs. This means that an appearance based method would most likely improve the performance of AU15 activation detection.

AU26 – jaw drop – is a very difficult action to detect for three reasons: firstly, AU26 and AU27 seem to differ only in the intensity of the jaw drop. In reality, the mechanisms behind the two AUs are quite distinct: AU26 is primarily caused by relaxation of the masseter muscle, while tensing the digastric muscle is the main cause for the activation of AU27 by pulling the jawbone down. Unfortunately, by judging the geometric displacement of the point on the chin only, it is hard to tell when AU26 stops and AU27 starts. The second difficulty with AU26 is that very often it is caused by a very small relaxation of the masseter, causing only a very small downward movement of the jawbone. More often than not, this occurs *withouth* parting of the lips. Thirdly, the point on the chin is not extremely well defined, especially on subjects with a lot of fat in their face. This makes the measurement of the displacement of the chin point difficult and this directly influences the recognition results of AU26.

One has to keep in mind that these results are for frame-by-frame AU activation detection. The manual AU labelling is very meticulous, even the slightest change in expression triggers the coding of an AU activation event by a human coder. Often, these expression changes start with a deformation of the local appearance of a facial feature rather than the movement of the facial feature. It is for this reason that our system always misses a number of frames in the beginning of a facial action and, for the same reason, at the end of the facial action. We will see in section 7.3, where we will perform AU event detection, that results actually look better if we count how many AUs were correctly detected on a per-video basis.

Frame-based evaluation on Cohn-Kanade database

This third study is performed to enable comparisons with other systems reported in the literature. Many older works were evaluated on the Cohn-Kanade Database. The Cohn-Kanade database was recorded for the purpose of basic emotion analysis and therefore only contains a limited number of AUs, with strong AU co-occurrence correlations. This is because the subjects were instructed to produce prototypic expressions of basic emotions (see section 3.3).

From this database, we created a dataset that can be used both for AU analysis and for emotion recognition. Image sequences were included in this set if two expert coders could reach consensus on the emotion shown. This was sometimes difficult as not all subjects were able to control their face

AU	Examples	Frames	Cl. Rate	Recall	Precision	F1-measure
1	68	883	0.918	0.808	0.844	0.826
2	50	657	0.939	0.791	0.879	0.833
4	54	857	0.870	0.604	0.658	0.630
5	37	421	0.904	0.566	0.629	0.596
6	39	535	0.930	0.789	0.811	0.800
7	31	415	0.870	0.268	0.315	0.290
9	30	357	0.928	0.676	0.497	0.573
10	26	302	0.914	0.403	0.401	0.402
12	42	727	0.930	0.827	0.844	0.836
15	19	264	0.969	0.500	0.283	0.361
20	34	381	0.908	0.466	0.582	0.517
24	17	297	0.935	0.395	0.497	0.440
25	19	1572	0.851	0.717	0.782	0.748
26	27	344	0.902	0.336	0.380	0.357
27	30	800	0.964	0.836	0.873	0.854
45	23	1243	0.943	0.584	0.408	0.480
Avg:			0.917	0.598	0.605	0.596

Table 6.4: Subject independent cross validation results for AU activation detection per frame on 153 examples from the Cohn-Kanade Database

as desired, making some expressions look like a mix of two or more emotions, for instance, a mix of happiness and surprise. AUs were again included in the set of target classes if there were at least 15 videos in which they occurred. As the second column of table 6.4 shows, due to the nature of this database, far fewer different AUs were present in the database recordings but the AUs that were present occurred in great abundance.

Results were measured in terms of c_r , r_r , p_r and ϕ_1 and are shown in table 6.4. As we can see, the results for this database are better (an increase in ϕ_1 of 6.6%). This may be due to two issues. Either the AUs are easier to detect due to the limited variation in the way expressions are displayed, or it is due to the higher number of examples that are available for each AU. To investigate the effect of the second issue, we carried out the experiment described in the next section.

Sample size effect

To find out what the effect the number of examples in a dataset has on the performance, we carried out an experiment. For AU1, we started with a subset of the data containing the minimum number of examples containing AU1 to perform 10-fold cross-validation on the subset (i.e. 10 examples for 10-fold cross-validation). Then we increased the size of the set by one positive example, and recorded the performance on this increased set. This was continued until all examples of AU1 were used. We

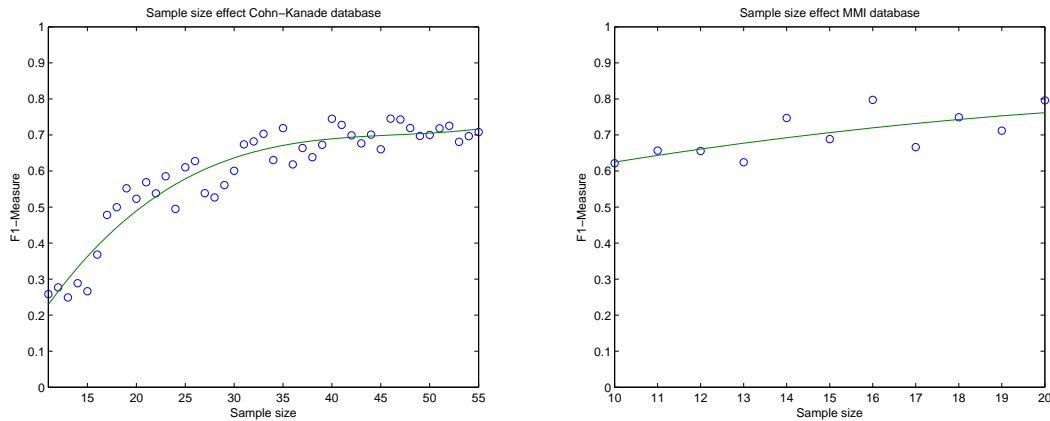


Figure 6.7: Evaluation of the effect of the number of positive examples in a dataset.

included all negative instances in each subset. This experiment was carried out on the data from both the MMI-Facial Expression Database and the Cohn-Kanade database.

The results, shown in figure 6.7, seem to indicate that indeed the sample size is of great importance. For the Cohn-Kanade Database, a minimum of 35 positive examples seems necessary. After 35 samples, the performance remains more or less stable. For the MMI facial expression database it is not possible to make any strong claims, due to the limited number of tests that we could run. Still, the figure seems to indicate that performance would increase if we would add more positive examples.

Based on these results, we can conclude that the higher recognition accuracy for AU-activation recognition on the Cohn-Kanade database is at least partially due to a larger sample size. It still does not tell us what effect the low expression variability has, but we will learn more about that in the next section.

Cross-database evaluation

A cross-validation study on data from a single database might have very good results, but it does not guarantee that the evaluated system performs well on truly novel data. In a sense, the only way to test whether a system performs well in the real world is to put it out there. However, this goes beyond our means. What we *can* do is test the generalisability of a system by training it on data from one database and test it on data from a second database. Both databases must be recorded completely independently of each other. That exactly is the case for the MMI-Facial Expression Database and the Cohn-Kanade Database.

AU	Frame-based		Event-based	
	Train MMI	Train Cohn-Kanade	Train MMI	Train Cohn-Kanade
1	0.454	0.368	0.661	0.255
2	0.433	0.389	0.762	0.467
4	0.421	0.411	0.541	0.414
5	0.366	0.262	0.447	0.149
6	0.475	0.592	0.429	0.571
7	0.073	0.032	0.129	0.211
9	0.472	0.552	0.495	0.286
10	0.175	0.114	0.232	0.109
12	0.190	0.499	0.635	0.400
15	0.248	0.180	0.372	0.229
20	0.240	0.323	0.277	0.341
24	0.182	0.198	0.333	0.292
25	0.770	0.646	0.799	0.746
26	0.278	0.155	0.293	0.203
27	0.398	0.546	0.589	0.591
45	0.308	0.308	0.442	0.622
Avg:	0.343	0.348	0.465	0.368

Table 6.5: F1-measure for cross-database AU detection per frame (second and third column) and per video (last two columns). The system was either trained on 244 examples from the MMI-Facial Expression Database and tested on 153 examples from the Cohn-Kanade Database or trained on Cohn-Kanade and tested on MMI.

The MMI-Facial Expression Database and the Cohn-Kanade database were recorded for different purposes. The MMI database was recorded for the purpose of research on atomic facial actions and includes, therefore, displays of singular and combined AU activations for all existing AUs³. The Cohn-Kanade database on the other hand was recorded to perform research on emotion recognition and no effort was made to capture singular AU activations of all AUs. As such, the AU distributions differ widely between the two databases and we could only test on AUs that were present in sufficient numbers in both sets.

We performed two tests: In the first test, we train the AU detector on all data from the MMI database and test it on data from the Cohn-Kanade database. Vice versa, in the second test we train on the Cohn-Kanade database and test on the MMI database. The results, measured in ϕ_1 , are shown in table 6.5.

The results in table 6.5 are grouped in two: a set of frame-based results and a group of event-based results. The latter determines whether an AU was activated during an entire video. In this case, a video was considered to contain an AU if at least one frame was detected as active. This is a good

³Note that though all AUs appear in the database, though not all possible AU combinations. This would have been a great feat indeed!

enough criterium for now, just to compare the two systems. Further on in this thesis we will explore other, more successful event detection methods.

The difference between frame-based and event-based results is important here, because the two databases are so different. The Cohn-Kanade database contains almost only neutral frames and frames in which the expression is going from neutral to its peak and hardly any in which the expression actually is at its peak. The MMI database on the other hand contains a well balanced number of all possible temporal phases of a facial action. To wit, the Cohn-Kanade data has a much lower ratio of high-intensity vs low-intensity frames. As the low-intensity AU activations are harder to detect than high-intensity AU activations, an AU detector will have a lower frame-based detection accuracy compared to a dataset where the ratio of high- vs. low-intensity frames is more balanced. This is why the frame-based results for the system trained on the Cohn-Kanade data seems so high, compared to that trained on MMI data. First of all, there are more frames with a high intensity in the MMI database to be detected by the Cohn-Kanade system, resulting in a relatively high performance. On the other hand, there are few high-intensity frames in the Cohn-Kanade data, resulting in a relatively low performance on frame-level for the MMI-system.

The results are more comparable at the event-based level. When we examine these, we see that the performance of the MMI-trained system is almost 10% higher than that of the Cohn-Kanade trained system. We believe that this is due to the low variability of facial expressions in the latter database. The results are even graver when we consider the effect of the sample size, examined in section 6.4.1. This should give the Cohn-Kanade system an edge over the MMI-trained system. Instead, it performs worse, probably because it expects AUs to be produced in combination with a number of other AUs, resulting in low generalisation to novel expressions. On the other hand, we see that the MMI-trained system generalises very reasonably on data from a completely different database.

This answers our question posed in section 6.4.1. It is both due to the large sample size *and* due to the limited variation in the way the emotive expressions are displayed that AU activation detection results are higher for the cross-validation study on the Cohn-Kanade dataset. In effect, it means that the Cohn-Kanade database is an easier database to benchmark on, and reported results obtained on the Cohn-Kanade database should be viewed as less remarkable as the same results obtained on the MMI database. Of course, it would be very hard to define exactly how much easier the Cohn-Kanade database is.

6.4.2 Event coding using AU frame activation information

The SVM classifier described above detects AUs per frame. Besides AU coding per frame, we also want to be able to perform so-called event coding. This means determining which AUs have been active at least once in a certain period of time. In our case, we want to determine which AUs were active in an image sequence.

The simplest way to perform event detection is to use a threshold on the number of frames predicted active by the frame based AU detector. As the SVM classifier adds an *a priori* unknown amount of noise to its output in the form of false positives and false negatives, fixing a threshold based on for example the minimal duration of an AU as observed by psychologists will not necessarily achieve optimal results. To overcome this problem we add a decision layer that will empirically learn a threshold θ based on the AUs automatically detected per frame. This frame activation detection output already contains the misclassification noise. The adaptive threshold we learn will use this knowledge to attain a higher probability of correctly performing the event coding on new test samples than a fixed threshold defined by a human expert, as it has knowledge about the errors typically made by the frame-based detector. The threshold is learned as follows:

Suppose the SVM determined that a test sample video \mathbf{x} has m frames where a certain AU A is present. Let \mathbf{N}_p be a vector containing for every video in which A is present the number of frames that the SVM predicted to have that AU present. Similarly, let \mathbf{N}_n be the vector containing per video in which A is *not* present the number of active frames that the SVM falsely predicted A to be present. Let m_p be the length of the shortest video segment belonging to \mathbf{N}_p , i.e. the video with the smallest number of true positives:

$$m_p = \min(\mathbf{N}_p) \quad (6.9)$$

Now let us filter the vector \mathbf{N}_n , leaving those elements smaller than m_p unchanged but setting all other values equal to zero, and denote this vector by \mathbf{N}_{np} . Now we find m_{np} as

$$0 \leq m_{np} = \max(\mathbf{N}_{np}) < m_p \quad (6.10)$$

which is the length of the longest segment of false positives in the subset of videos \mathbf{N}_{np} that have in

total fewer false positives than the smallest number of true positives in any of the videos, N_p . Finally, let us denote the number of active frames in a test sequence by m . The threshold θ that is used to decide whether the test sample \mathbf{x} contains A (i.e. $m > \theta$) is now defined as:

$$\theta = \frac{m_p + m_{np}}{2} \quad (6.11)$$

In effect what we have done here is maximising the margin between the true positives and the false positives.

The event detection results using equation 6.11 on the frame-based AU activation detection output are shown in table 6.6 for data from the MMI database and in table 6.7 for data from the Cohn-Kanade database. As we can see, the results are not much better than the frame-based results. While we have improved the precision rates by recovering from a number of false positives, we have at the same time decreased the recall by creating more false negatives. In the next section, we will propose to analyse the temporal dynamics of an AU each time that AU was predicted active in one or more frames of an image sequence. We will see that when we do this, AU event detection results improve significantly.

AU	Examples	Cl. Rate	Recall	Precision	F1-measure	$\theta = 0$
1	22	0.909	0.955	0.500	0.656	0.618
2	25	0.881	0.880	0.458	0.603	0.597
4	38	0.872	0.763	0.569	0.652	0.544
5	19	0.885	0.579	0.355	0.440	0.492
6	27	0.885	0.852	0.489	0.622	0.537
7	15	0.905	0.667	0.357	0.465	0.435
9	15	0.909	0.867	0.394	0.542	0.489
10	17	0.872	0.471	0.267	0.340	0.426
12	17	0.856	0.765	0.295	0.426	0.444
13	14	0.909	0.857	0.375	0.522	0.510
15	15	0.885	0.600	0.290	0.391	0.417
16	18	0.807	0.389	0.163	0.230	0.346
18	16	0.885	0.625	0.312	0.417	0.375
20	15	0.905	0.933	0.389	0.549	0.583
22	15	0.868	0.867	0.302	0.448	0.417
24	15	0.885	0.800	0.324	0.462	0.417
25	105	0.823	0.933	0.729	0.819	0.824
26	32	0.704	0.548	0.227	0.321	0.362
27	15	0.934	1.000	0.484	0.652	0.437
30	15	0.897	0.857	0.343	0.490	0.292
43	15	0.909	0.867	0.394	0.542	0.510
45	107	0.877	0.916	0.824	0.867	0.846
Avg:		0.876	0.772	0.402	0.521	0.519

Table 6.6: Subject independent cross validation results for AU event detection using adaptive thresholds on 244 examples from the MMI Facial Expression Database. The last column shows the results for fixing $\theta = 0$.

AU	Examples	Cl. Rate	Recall	Precision	F1-measure	$\theta = 0$
1	68	0.915	0.864	0.911	0.887	0.780
2	50	0.935	0.929	0.848	0.886	0.851
4	54	0.856	0.886	0.633	0.738	0.644
5	37	0.889	0.818	0.711	0.761	0.763
6	39	0.922	0.912	0.775	0.838	0.827
7	31	0.784	0.522	0.353	0.421	0.351
9	30	0.922	0.800	0.667	0.727	0.686
10	26	0.830	0.650	0.406	0.500	0.444
12	42	0.954	0.972	0.854	0.909	0.742
15	19	0.928	0.667	0.167	0.267	0.438
20	34	0.850	0.586	0.607	0.596	0.667
24	17	0.948	0.636	0.636	0.636	0.563
25	19	0.830	0.881	0.766	0.819	0.897
26	27	0.843	0.474	0.391	0.429	0.328
27	30	0.987	1.000	0.926	0.962	0.750
45	23	0.922	0.778	0.636	0.700	0.723
Avg:		0.895	0.773	0.643	0.692	0.653

Table 6.7: Subject independent cross validation results for AU activation event detection using adaptive thresholds on 153 examples from the Cohn-Kanade Database. The last column shows the results for fixing $\theta = 0$.

Chapter 7

Action Unit temporal analysis

In the previous chapter, we have investigated the problem of analysing the morphology of a facial expression. That is, we have identified *which* facial muscles (i.e. AUs) were used. What we haven't investigated is *how* those facial muscles were used. In this section we will go deeper into this issue by presenting methods that can recognise exactly when a facial action starts, when it reaches a peak, when a peak ends, and when it has returned to a neutral state. As a by-product of this temporal phase analysis, we will also be able to tell how many peaks there were, how long the overall facial action lasted and how fast the facial action reached its peak. This information will be used later in chapter 9 to distinguish posed from spontaneous expressions.

A facial action (AU) can be in any of four possible temporal phases: the onset phase, the apex phase, the offset phase and the neutral phase. The onset phase is defined as the period in which the visible change in appearance and/or displacement of facial features caused by the contraction of the related facial muscle(s) is growing stronger. For instance, in the onset phase of AU1, which causes the inner portion of the eyebrows to be raised, the inner eyebrow will move up, away from its initial neutral position. In the apex phase of a facial action, the corresponding muscle(s) are contracted, but the contraction does not increase anymore, nor does it decrease. In turn, the related facial points, e.g. the points on the inner eye-brows, remain in a fixed position. While the apex phase indicates that a peak in muscle contraction has been reached, the first apex phase reached need not be the only peak, nor does it have to be the peak with maximum intensity. Multiple apices of varying intensities may occur before the facial muscle returns to its neutral position. During the offset phase the muscles relax again, effectively causing the related facial points to move towards their position typical for an

expressionless face. When a facial action has returned to the neutral phase, we regard that action to be finished. If after some time another onset phase begins, we regard this to be a new facial action.

In some cases the activation of an AU involves relaxation of certain muscles instead of contraction of these muscles. In such cases one should interchange ‘contraction’ with ‘relaxation’ in the explanation above. Often the order of the temporal segments is neutral-onset-apex-offset-neutral, but more complex combinations such as neutral-onset-apex-onset-apex-offset-neutral (i.e. multiple-apex) facial actions are possible as well.

In this chapter we will propose and compare three methods to identify the temporal phases of an AU. We will show that the proposed combination of Support Vector Machine-Hidden Markov Model performs extremely well. We will also show that using the output of the temporal analysis system for event detection greatly improves event detection accuracy.

7.1 Temporal analysis per frame

As a facial action can at any moment in time be in only one of the four temporal segments, we consider the problem of AU temporal segment recognition to be a four-valued multiclass classification problem. Initially, we tried a more or less brute-force approach, that does not model the evolution of a facial action explicitly. In this approach, we determine for each frame separately in what temporal phase the facial action is at that moment, independently from all other frames in a video.

To solve this problem we used a one-vs-one approach to multiclass SVMs (mc-SVMs). In this approach, for each AU and every pair of temporal segments we train a separate sub-classifier specialised in the discrimination between two temporal segments. This results in $\sum_{i=1}^{K-1} i = 6$ sub-classifiers that need to be trained per AU ($C = \{1, 2, 3, 4\}$, and $K = |C| = 4$ being the number of classes in our multiclass problem). When a new test example $x(t)$ is introduced to the mc-SVM, every sub-classifier returns a prediction of the class $c(t) \in C$, and a majority vote is cast to determine the final output $c(t)$ of the mc-SVM for an example at time t .

As we use the same features for the recognition of temporal segments that were used for the detection of AU activation and the ratio of positive/negative samples is very similar to that of the AU activation detection problem, we again apply GentleBoost feature selection. GentleBoost determines which features will be used for training and testing the sub-classifiers. Table 7.5 lists which feature was

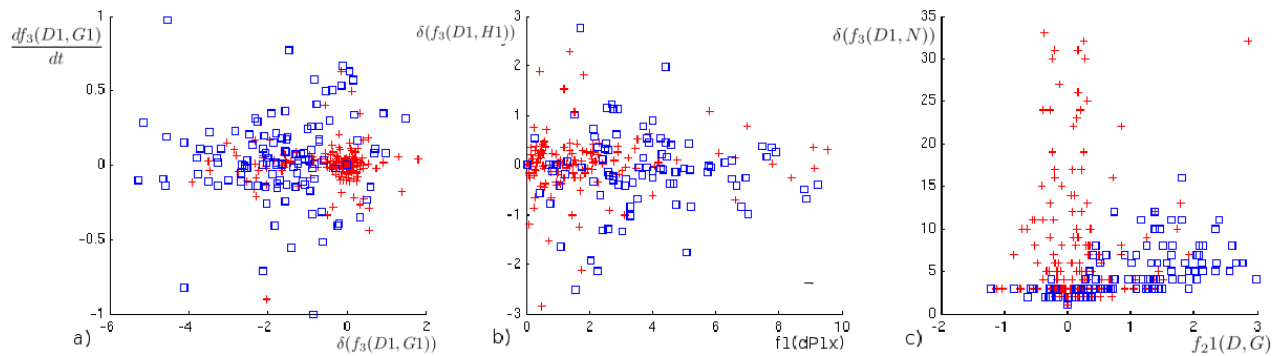


Figure 7.1: Scatter plots of pairs of the most informative mid-level feature parameters selected by GentleBoost for: (left) the onset temporal segments of AU1, (mid) the apex temporal segments of AU1 and (right) the offset temporal segments of AU1. Crosses denote spontaneous brow data while squares denote data of deliberately displayed brow actions.

selected to be the most important for each temporal segment of every AU. The most informative features for classification of onset vs. neutral, apex vs. neutral, and offset vs. neutral temporal segments of AU1 are depicted in Fig. 7.1. A more detailed investigation on the characteristics of the features regarding AU temporal dynamics is given in section 7.4.

A different sub-classifier is trained for each combination of classes. As the task of these sub-classifiers is to distinguish between two classes only, we train each sub-classifier with a different subset of the training data that contains only examples that are either of the positive or of the negative class of that sub-classifier. For instance, if we learn an onset vs. offset classifier, we remove all frames labelled neutral or apex from the training data. Similarly, the feature selection is performed on this subset of data, resulting in a different set of selected features for each sub-classifier.

To investigate how well this approach to the analysis of facial expressions works, we performed the following evaluation study. In our fully automatic system, we intend to apply the temporal analysis of an AU only on videos in which that AU was found to be active. At this point we want to focus on the performance of the temporal segment recognition element, separately from the performance of the AU activation detection element (which was evaluated thoroughly in chapter 6). Therefore, we assume a perfect AU activation detection, i.e. we use the manual labelling of AUs to decide whether an AU was active in a video or not.

Because after this assumption we know whether a video contains the AU, and we only need to distinguish between the temporal segments of that particular AU, it is sufficient to use as training data only the videos in which that AU is active. We again perform leave-one-subject-out cross validation

AU	Neutral	Onset	Apex	Offset
1	0.787	0.615	0.506	0.434
2	0.724	0.577	0.651	0.447
4	0.687	0.525	0.668	0.260
5	0.698	0.275	0.440	0.234
6	0.751	0.388	0.402	0.316
7	0.616	0.114	0.300	0.120
9	0.835	0.594	0.833	0.163
10	0.805	0.572	0.726	0.382
12	0.911	0.679	0.717	0.622
13	0.924	0.802	0.691	0.616
15	0.749	0.264	0.733	0.215
16	0.763	0.470	0.578	0.450
18	0.919	0.681	0.680	0.487
20	0.832	0.596	0.896	0.464
22	0.847	0.675	0.429	0.337
24	0.465	0.263	0.433	0.029
25	0.839	0.612	0.774	0.563
26	0.764	0.510	0.644	0.382
27	0.734	0.669	0.506	0.659
30	0.705	0.272	0.433	0.376
43	0.880	0.541	0.595	0.639
45	0.963	0.717	0.516	0.629
46	0.679	0.243	0.258	0.225
Avg:	0.777	0.507	0.583	0.393

Table 7.1: Classification accuracy of multiclass gentleSvm at distinguishing the four temporal phases neutral, onset, apex and offset, measured per frame in terms of F1-measure.

on this subset of the data. We have evaluated our proposed methods on the 244 videos from the MMI database only, because the videos in the MMI-Facial Expression Database display the full neutral-expressive-neutral pattern and the videos from the Cohn-Kanade database stop once they reach their peak. It is essential to have data with this complete neutral-expressive-neutral pattern, as it is exactly this pattern of facial actions that we are interested in. Classification accuracy is measured per frame. The results of this study using mc-SVMs are shown in table 7.1.

7.2 Temporal model using temporal dynamics grammatical rules

The frame-based onset, apex and offset detector described in section 7.1 is not perfect. For one thing, it lacks the ability to model the relationship between consecutive frames. Let us think of a facial expression as being a sentence, with the AU activations acting as the words and an AU’s temporal segments as the letters of a word. To improve the analysis of facial action temporal dynamics, one fact

to consider is that like a sentence in natural language, a facial action should obey certain grammatical rules. For instance, an AU activation cannot begin with an apex segment. Even though the preceding onset or offset phase may be short, it always precedes the apex phase of an AU activation. Similarly, the transition from an apex phase to a neutral phase should involve an offset phase, though the temporal duration of this offset may be (very) short. In our system, we implemented these grammatical rules of an AU activation by means of two temporal filters.

The first filter enforces a minimal duration of a temporal segment and effectively removes all isolated misclassifications from the mc-SVM prediction. We observed from our data that no temporal segment has a duration shorter than 3 frames, which is in our case $\frac{1}{8}$ th of a second. To decide what the temporal segment of the prediction c_t will be, we apply a sliding window filter, casting a majority vote on the predictions $[c_{t-2} \dots c_t \dots c_{t+2}]$. In case of a tie, we decide to assign the value of c_{t+1} to prediction c_t .

Next we apply a filter that enforces certain temporal segment transitions. There are two violations of grammatical rules that we check for.

1. Neutral to Apex. If at time t a transition is made from the neutral phase to the apex phase, we change the values of the predictions $[c_t \dots c_{t+2}]$ to ‘onset’, effectively inserting a minimal length onset phase.
2. Apex to Neutral. If at time t a transition is made from the apex phase to the neutral phase, we change the values of the predictions $[c_t \dots c_{t+2}]$ to ‘offset’, effectively inserting a minimal length offset phase.

To test the effect of these grammatical rules, we used the same evaluation procedure as proposed in section 7.1. The results for the frame-based recognition using grammatical rule post-processing of the mc-SVM output are shown in table 7.2.

7.3 Temporal model using Hidden Markov Models

Traditionally, HMMs have been used with great effect to model time in classification problems. But while the evolution of a facial action over time can be represented very efficiently by HMMs, the frame-by-frame emission probabilities of the various states are normally modelled by fitting Gaussian mixtures on the features. These Gaussian mixtures are fitted using likelihood maximisation, which

AU	Neutral	Onset	Apex	Offset
1	0.826	0.598	0.483	0.488
2	0.737	0.570	0.645	0.492
4	0.681	0.414	0.553	0.284
5	0.734	0.259	0.455	0.310
6	0.748	0.349	0.400	0.303
7	0.813	0.171	0.311	0.286
9	0.855	0.669	0.849	0.492
10	0.821	0.552	0.681	0.386
12	0.915	0.655	0.701	0.638
13	0.921	0.804	0.715	0.642
15	0.783	0.285	0.716	0.487
16	0.768	0.453	0.595	0.433
18	0.924	0.666	0.667	0.489
20	0.851	0.610	0.904	0.535
22	0.861	0.690	0.467	0.389
24	0.506	0.307	0.422	0.142
25	0.845	0.621	0.786	0.576
26	0.778	0.511	0.630	0.380
27	0.782	0.674	0.516	0.710
30	0.721	0.285	0.409	0.419
43	0.891	0.523	0.614	0.690
45	0.959	0.484	0.151	0.606
46	0.679	0.333	0.279	0.310
Avg:	0.800	0.499	0.563	0.456
Avg:	0.800	0.499	0.563	0.456

Table 7.2: Classification accuracy of multiclass gentleSvm classifier followed by a filtering process using expression grammatical rules for distinguishing the four temporal phases neutral, onset, apex and offset, measured per frame, in terms of F1-measure.

assumes correctness of the models (i.e. the feature values should follow a Gaussian distribution) and thus suffers from poor discrimination [Bourlard and Morgan, 1998]. Moreover, it results in mixtures trained to model each class and not to discriminate one class from the other.

SVMs on the other hand are incapable of modelling time, but they discriminate extremely well between classes. Using them as emission probabilities might very well result in an improved recognition. We therefore train one SVM for every combination of classes (i.e., temporal phases neutral, onset, apex, and offset), just as described in section 7.1. We then use the output of the mc-SVMs to compute the emission probabilities. This way we effectively have a hybrid SVM-HMM system. This approach has been previously applied with success to speech recognition [Kruger et al., 2005].

HMMs work in a probabilistic framework. Unfortunately we cannot use the output of a SVM directly as a probability measure. The (unsigned) output $h(\mathbf{x})$ of a SVM is a distance measure between a test pattern and the separating hyper plane defined by the support vectors. There is no clear relationship with the posterior class probability $p(y = +1|\mathbf{x})$ that the pattern \mathbf{x} belongs to the class $y = +1$. However, Platt proposed an estimate for this probability by fitting the SVM output $h(\mathbf{x})$ with a sigmoid function [Platt, 2000]:

$$p(y = +1|\mathbf{x}) = g(h(\mathbf{x}), A, B) \equiv \frac{1}{1 + \exp(Ah(\mathbf{x}) + B)} \quad (7.1)$$

The parameters A and B of eq.(7.1) are found using maximum likelihood estimation from a training set p_i, Y_i with $p_i = g(f(\mathbf{x}_i), A, B)$ and target probabilities $Y_i = (y_i + 1)/2$. The training set can but does not have to be the same set as used for training the SVM.

Since SVMs are *binary* classifiers we use a one-versus-one approach to come to a multiclass classifier. This approach is to be preferred over the one-versus-rest approach as it aims to learn the solution to a more specific problem, namely, distinguishing between one class from one other class at a time. This is in line with our idea to use SVMs for discrimination specificity and HMMs to model time. For this pairwise classification we need to train $K(K - 1)/2$ SVMs, where in our case $K = 4$ is the number of temporal phases.

Our HMM consists of four states, one for each temporal phase. For each SVM we get, using Platt's method, pairwise class probabilities $\mu_{ij} \equiv p(c_i|c_i \text{ or } c_j, \mathbf{x})$ of the class (HMM state) c_i given the feature vector \mathbf{x} and that \mathbf{x} belongs to either c_i or c_j . These pairwise probabilities are transformed

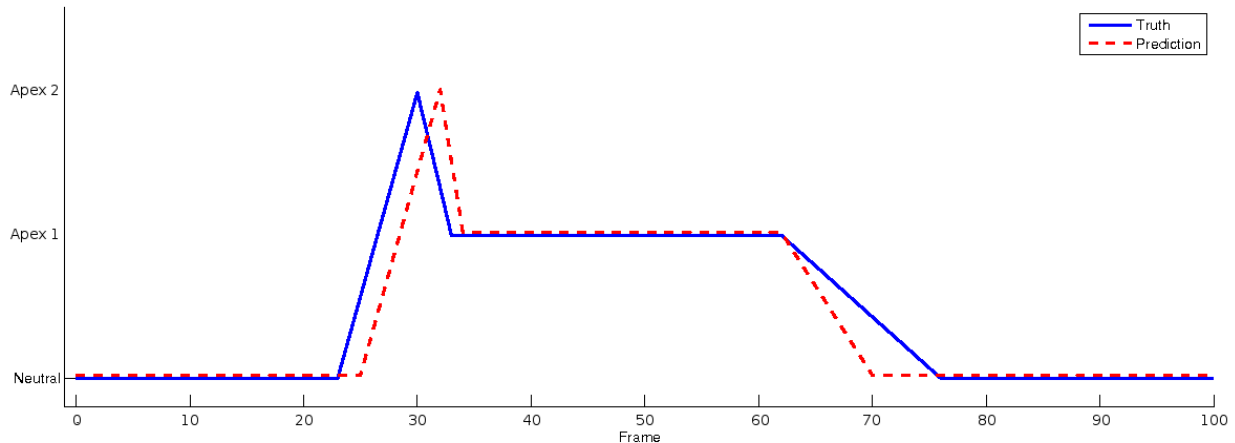


Figure 7.2: An example of AU temporal phase recognition for AU25. The solid line shows the true values of the phase labelling per frame and the dotted line the prediction by the SVM-HMM. Horizontal lines depict either a neutral or an apex phase, an upward slope signifies an onset phase and a downward slope an offset phase.

into posterior probabilities $p(c_i|\mathbf{x})$ by

$$p(c_i|\mathbf{x}) = 1 / \left[\sum_{j=1, j \neq i}^C \frac{1}{\mu_{ij}} - (K - 2) \right] \quad (7.2)$$

Finally, the posteriors $p(c_i|\mathbf{x})$ have to be transformed into *emission probabilities* by using Bayes' rule

$$p(\mathbf{x}|c_i) \propto \frac{p(c_i|\mathbf{x})}{p(c_i)} \quad (7.3)$$

where the a-priori probability $p(c_i)$ of class c_i is estimated by the relative frequency of the class in the training data.

We again evaluated our proposed hybrid SVM-HMM method on the 244 videos from the MMI database. The results for the frame-based recognition results are shown in table 7.3. Figure 7.2 shows one example of the recognition of the temporal phases of a video containing an AU 25 activation. The figure shows that the prediction (red line) is one frame late at predicting the first and second apex phases. It also predicts the last offset phase to stop 6 frames too early. The SVM-HMM system did recognise correctly that there are two apex phases.

Figure 7.3 shows the relative increase of performance in recognising each temporal segment, comparing the three methods: using mcSVMs only, using mcSVMs with grammatical rule post-processing and

AU	Neutral	Onset	Apex	Offset
1	0.790	0.669	0.585	0.536
2	0.848	0.544	0.730	0.642
4	0.690	0.521	0.615	0.334
5	0.610	0.352	0.561	0.292
6	0.807	0.469	0.693	0.374
7	0.784	0.100	0.390	0.108
9	0.895	0.756	0.887	0.462
10	0.855	0.587	0.790	0.323
12	0.931	0.693	0.773	0.679
13	0.926	0.847	0.750	0.642
15	0.791	0.339	0.742	0.357
16	0.815	0.481	0.600	0.384
18	0.914	0.569	0.740	0.592
20	0.883	0.734	0.860	0.583
22	0.864	0.701	0.469	0.373
24	0.507	0.257	0.547	0.037
25	0.865	0.634	0.776	0.631
26	0.751	0.490	0.583	0.417
27	0.720	0.747	0.858	0.708
30	0.787	0.461	0.541	0.415
43	0.937	0.476	0.758	0.728
45	0.971	0.780	0.653	0.710
46	0.618	0.146	0.182	0.239
Avg:	0.807	0.537	0.656	0.459

Table 7.3: Classification accuracy of hybrid SVM-HMM classifier at distinguishing the four temporal phases neutral, onset, apex and offset, measured per frame, in terms of F1-measure.

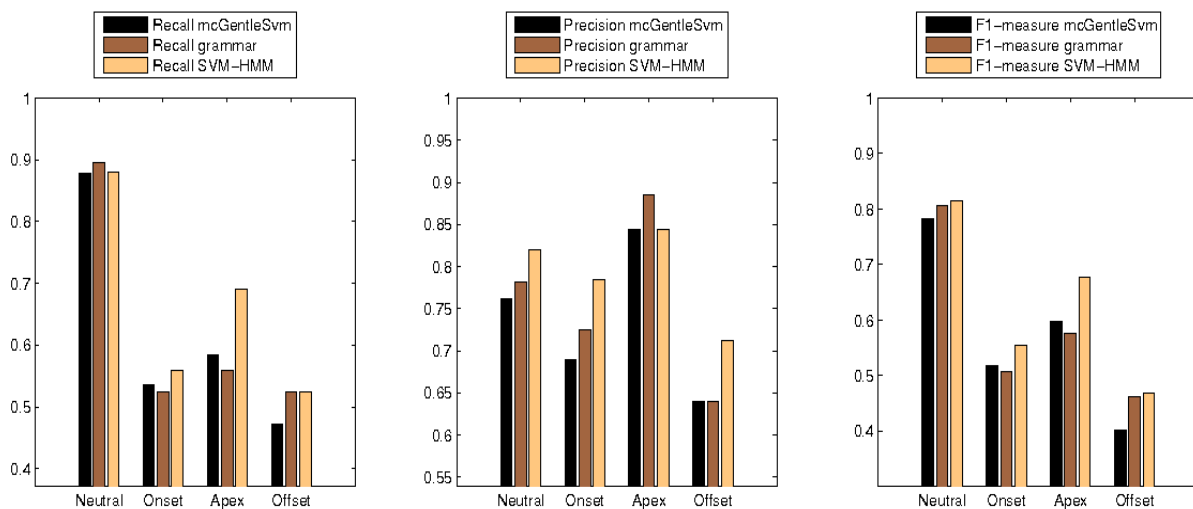


Figure 7.3: Comparison of the classification results shown per temporal phase (onset, apex, offset and neutral). The results shown are the average over all 22 AUs.

using the hybrid SVM-HMMs proposed in this section. In this figure we have averaged the results per temporal phase over all AUs.

When we take a closer look at the results of the detection of the temporal segments, we see that compared with the multiclass gentleSvm method, the detection of the apex phase has increased most from introducing the HMM. This is followed closely by the offset phase. The apex phase had an increase in ϕ_1 of 8%, the offset 6.8%, the onset phase 3.6% and the neutral phase 3.4% (relative to mc-SVMs). The fact that the neutral phase benefits least from the addition of the HMM is as expected, because by its very nature it is not a dynamic part of the facial action. The effect of applying the grammatical rules is less successful. While it attains good results for the offset phase and in a limited way for the neutral phase, it actually decreases the accuracy of the onset and apex phase recognition.

While the classification accuracy per frame is a good way to compare various approaches, it might not always provide the best insight to the overall performance of an AU temporal phase recognition system. For instance, think of the following extreme example: we have an onset starting at frame 2 and lasting 1 frame, followed by an apex phase of 1 frame and an offset phase of 1 frame, after which the AU returns to its neutral state. The entire AU thus lasted three frames, and could be encoded as $[0 \ 1 \ 2 \ -1 \ 0]$. Now let us assume that our temporal analysis system returned a classification of $[0 \ 0 \ 1 \ 2 \ -1]$, that is, the same result but shifted in time by one frame. This would score 0% recall, precision, classification rate and F1-measure for all phases except the neutral phase. Yet this would not be a

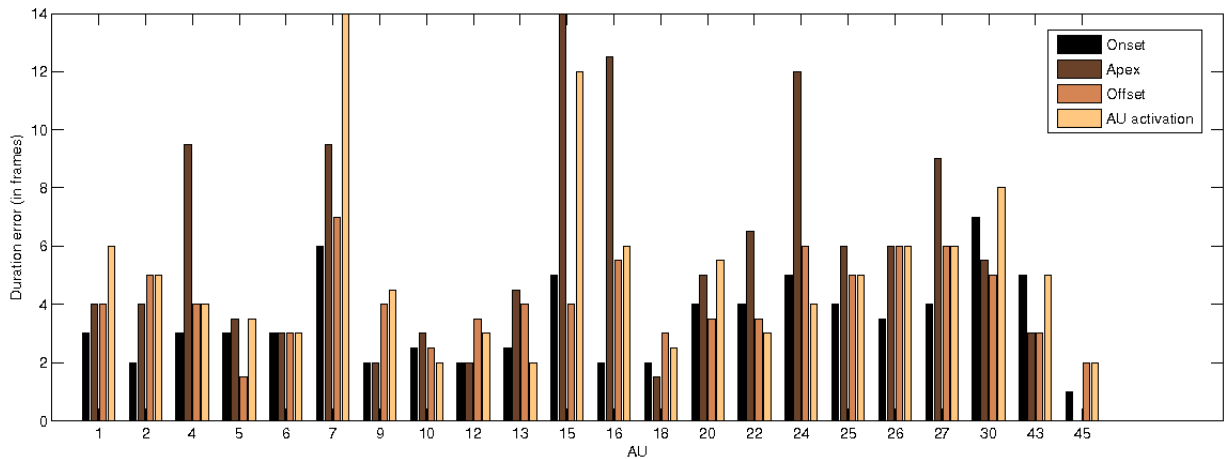


Figure 7.4: Phase duration error of the detected temporal phases onset, apex and offset, and the entire facial action. Results are averaged per AU, and measured frames.

fair judgement of the system. In essence, it performed really well, it only introduced a slight delay.

We therefore take a closer look at some other statistics of the performance of the AU temporal analysis system that employs the hybrid SVM-HMM classifier and GentleBoost feature selection. To de-correlate these results from the classification accuracy results, we only investigated the videos for which the SVM-HMM system predicted at least one frame to be non-neutral, enhancing the timing results in the process. First of all we looked into the durations of the facial actions, both the total duration of an AU (i.e. the number of consecutive frames that were predicted to be non-neutral) as well as the durations of the temporal phases separately. Figure 7.4 shows the statistics for this analysis. The duration error is measured in frames. The figure shows the average number of frames that a temporal phase duration or the entire AU activation duration is off, averaged per AU. We can see that for most AUs, the average error per temporal phase is less than 4 frames. The apex temporal phase has the largest error. We can also see from figure 7.4 that the error of the total AU activation duration is far less than the sum of the temporal phase duration errors. This is because usually, if the apex phase has been predicted to last too long, consequently the offset phase will start late and results in an error in the offset phase duration, which effectively gets double counted.

Sometimes it is not informative enough to look at the number of frames that the duration of a temporal phase or an entire facial action is mis-predicted. An error of four frames has much larger consequence for the onset phase, which typically lasts between 4 and 10 frames, than for the apex phase, which can last multiple seconds, i.e. almost hundreds of frames. We therefore present the temporal phase

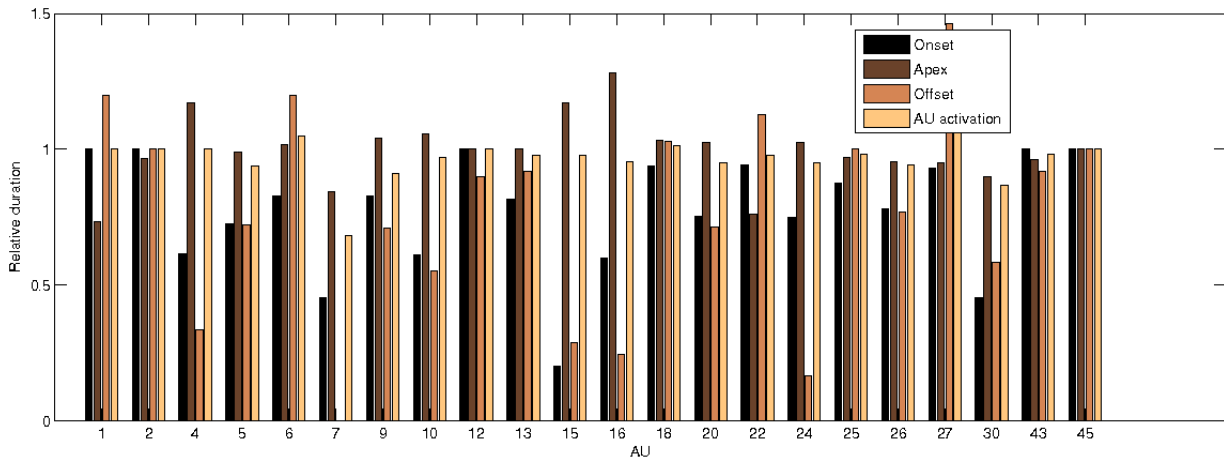


Figure 7.5: Relative phase duration of the detected temporal phases onset, apex and offset, and the entire facial action. Results are averaged per AU, and measured frames.

duration results in a slightly different way in figure 7.5. In that figure we show the relative duration of the temporal phases of the AU activation. That is, if a phase was manually labelled to last 4 frames and predicted to last 8 frames, figure 7.5 would show a value of 200% for that phase. Compared with figure 7.4 we notice now that using this measure the error of the apex and total AU activation duration are almost negligible. This is because their total durations are much longer.

Next we look into the timing of the temporal phase transitions. In figure 7.6 we show for the three most important transitions: neutral to onset, onset to apex, and apex to offset how many frames they were early (negative numbers) or late (positive numbers), on average. For most AUs, the timing is very accurate, with an average error of less than two frames early or late. The figure indicates that the system usually predicts the apex a little bit too early and the onset a little bit too late. For the offset the values differ per AU. This is also the temporal phase for which the most AUs result in a large (> 3 frames) timing error.

Figure 7.6 shows how many frames *on average* a transition is late or early. However, we would like to know a bit better exactly how often an AU is early or late, and how often it is exactly on time. This is shown in figure 7.7. The figure shows that a few temporal phases of some AUs are persistently predicted early (e.g. the offset of AU1, the apex of AU7 or the apex of AU15), but very often phases are predicted sometimes to start early and sometimes to start late. It is noteworthy that the onset, and therefore the start of the entire facial action, is almost never early. We believe this is because most facial actions start with only a change in appearance, followed shortly after by motion of the related

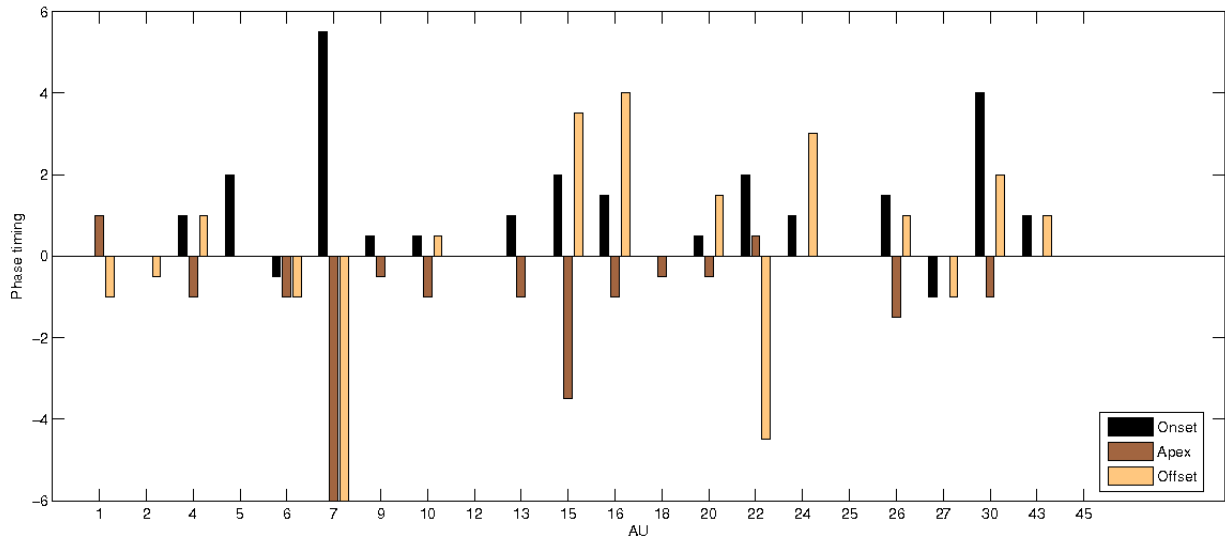


Figure 7.6: Average number of frames a phase starts late (positive values) or early (negative values), for the temporal phases onset, apex and offset, per AU.

facial points. As our features only encode information of the facial point motion, our classifiers cannot detect this very first change of appearance, which marks the beginning of the facial action. Only the AUs 2, 4, 9, 12, 18 and 45 start exactly on time for a significant number of predictions. Note that the detection of the temporal segments of AU45 is nearly perfect. This has two reasons: firstly, we have a very large number of examples for this AU, no less than 107 videos contained an AU45 activation. Secondly, AU45 is a purely reflexive action and the temporal pattern is extremely robust. Our system has completely captured the temporal dynamics of this action.

7.4 Feature set evaluation

At this point we want to revisit the analysis of the three feature sets proposed in section 5.3. In section 6.4.1 we analysed their importance relative to AU activation detection and found that F_D performed best, while F_S and F_W performed almost equally well. For the recognition of the temporal phases of AUs we have repeated the evaluation study performed in section 6.4.1, only now with the goal to recognise each frame as belonging to either the neutral, onset, apex or offset phases. Table 7.4 shows the results of this test.

As expected, the difference between the purely static features F_S and the local time parameter features F_W is quite distinct now. This is most visible for the highly dynamic phases onset and offset, where

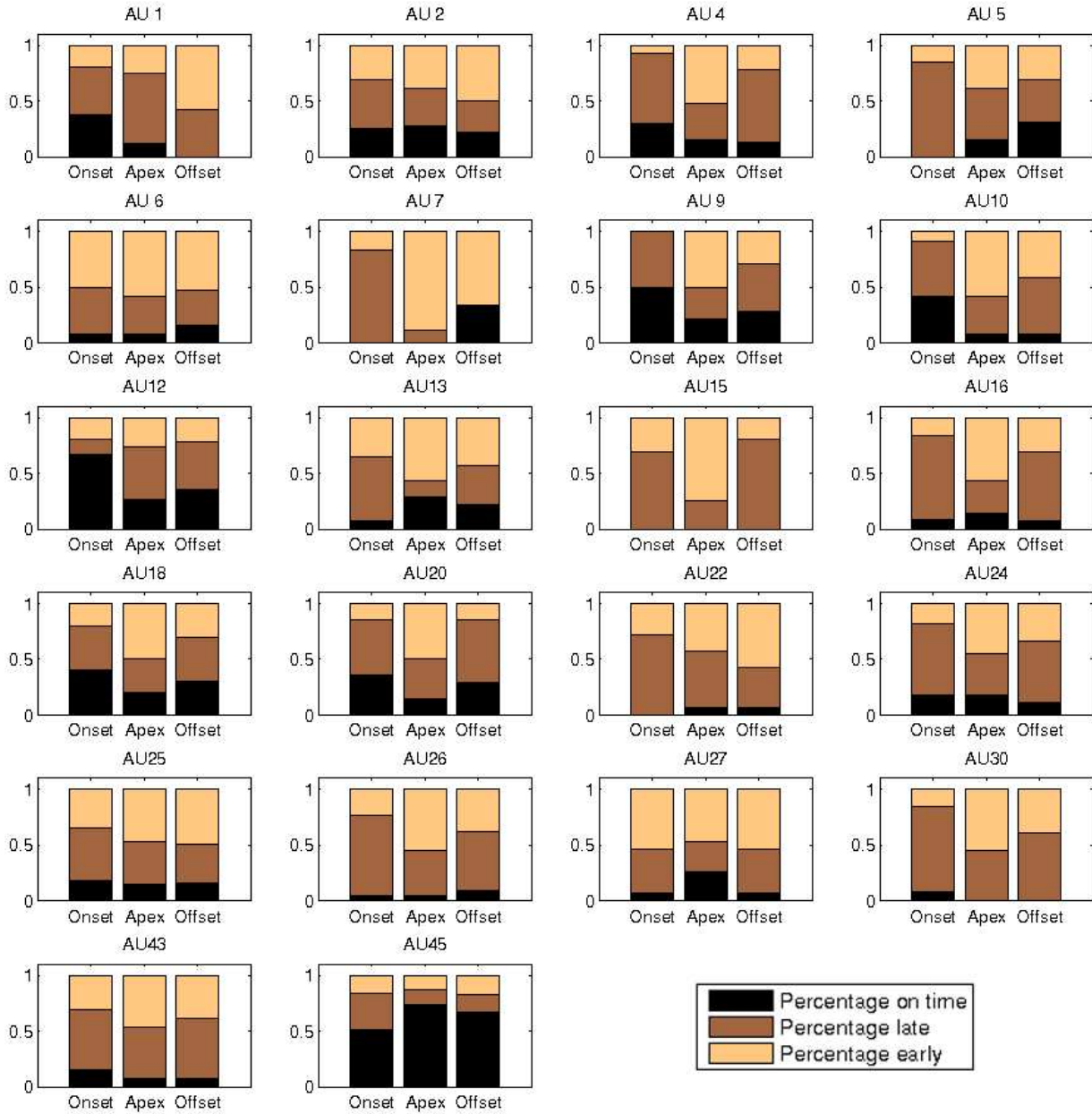


Figure 7.7: Breakdown of how often a temporal phase starts early, late, or on time.

Features	Avg. cl. rate	Avg. ϕ_1	ϕ_1 neutral	ϕ_1 offset	ϕ_1 onset	ϕ_1 apex
F_S	0.799	0.496	0.903	0.200	0.245	0.638
F_D	0.856	0.646	0.926	0.440	0.489	0.731
F_W	0.847	0.616	0.923	0.396	0.463	0.681
$\{F_S, F_D\}$	0.862	0.655	0.933	0.434	0.502	0.751
$\{F_S, F_W\}$	0.848	0.610	0.926	0.396	0.452	0.667
$\{F_D, F_W\}$	0.859	0.663	0.927	0.473	0.517	0.733
$\{F_S, F_D, F_W\}$	0.868	0.688	0.936	0.498	0.541	0.777

Table 7.4: Performance of various feature set combinations for the problem of recognising the temporal phases of Action Units. The fourth to seventh column show the average values over all AUs, per temporal phase. seventh column show the average values over all AUs, per temporal phase. The second and third columns show the average values over all temporal phases and all AUs.

the feature set F_W outperforms F_S by almost a factor two. Also, the addition of F_W to F_S and F_D is more effective at solving this problem of temporal dynamics analysis. F_S cannot encode any temporal dynamics information and is outright poor at recognising the onset and offset phases. F_D is once again the best performing feature set, but the combination of all features outperforms F_D by a good 5% ϕ_1 increase for the non-neutral phases.

We again have a closer look at exactly which features are chosen by GentleBoost for the recognition of the temporal phases of AUs. In our proposed AU temporal dynamics analysis method, the temporal phases are distinguished pairwise from each other. Table 7.5 lists the best feature selected to separate examples for each of the 6 combinations, as selected by GentleBoost. For brevity, the time parameter t is left out of the feature definitions, but it is understood that the features have a different value for each moment in time t (see section 5.3).

7.5 Event detection using AU temporal analysis

We can now use the output of any of the AU temporal analysis methods to perform event detection, instead of the adaptive threshold proposed in section 6.4.2. At this point we want to evaluate how well the entire fully automatic system performs at event detection, combining AU activation detection and AU temporal analysis. To detect an event, each video in which at least one frame has been determined to be active is passed on to the SVM-HMM temporal analysis sub-system described in this chapter. The analysis of the temporal dynamics might result in a neutral-expressive-neutral pattern, but the system might also decide that the entire sequence was neutral after all. If, after recognition of the temporal phases, any of the frames remains non-neutral, we determine that AU to be present within

AU	N vs Of	N vs On	N vs A	Of vs On	Of vs A	On vs A
1	$\delta(f_3(D1, N))$	$\frac{df_3(D1, G1)}{dt}$	$\delta(f_3(D1, H1))$	$\frac{df_3(D1, G1)}{dt}$	$\delta(f_3(D1, H))$	$\frac{df_2(D)}{dt}$
2	$\delta(f_3(D1, F))$	$\frac{df_3(E1, G1)}{dt}$	$\delta(f_3(A, D1))$	$\frac{df_3(E1, G1)}{dt}$	$\frac{df_3(E, G)}{dt}$	$\frac{df_3(E1, G1)}{dt}$
4	$f_{21}(B1, D)$	$\frac{df_4(D, G)}{dt}$	$\delta(f_3(D1, H1))$	$\frac{df_4(D, G)}{dt}$	$\delta(f_3(B1, D1))$	$f_{24}(B1, D)$
5	$\delta(f_3(F, G))$	$\delta(f_4(F1, G))$	$\delta(f_3(F, G))$	$\frac{df_3(F1, G1)}{dt}$	$\delta(f_3(G1, H1))$	$f_{21}(F, G)$
6	$f_{24}(K, J)$	$\frac{df_3(I, M)}{dt}$	$\delta(f_3(I, M))$	$\frac{df_3(I, M)}{dt}$	$\frac{df_3(J, M)}{dt}$	$f_{21}(J, M)$
7	$\delta(f_4(A1, G))$	$\delta(f_4(A, G))$	$\delta(f_3(B1, D1))$	$\frac{df_4(F1, G)}{dt}$	$\delta(f_3(B1, E1))$	$\delta(f_3(B1, D1))$
9	$\delta(f_3(A1, M))$	$\frac{df_3(K, M)}{dt}$	$\delta(f_3(G, M))$	$\frac{df_3(L, M)}{dt}$	$\frac{df_3(K, M)}{dt}$	$\delta(f_3(A1, M))$
10	$\delta(f_3(K, M))$	$\delta(f_3(K, M))$	$\delta(f_3(K, M))$	$\frac{df_3(K, L)}{dt}$	$\delta(f_3(K, M))$	$f_{21}(K, L)$
12	$\delta(f_3(K, J))$	$\delta(f_3(J, L))$	$\delta(f_3(I, J))$	$\frac{df_4(I, N)}{dt}$	$\delta(f_4(B1, I))$	$f_{24}(F1, J)$
13	$f_{24}(H, J)$	$\frac{df_3(I, M)}{dt}$	$\delta(f_3(J, M))$	$\frac{df_2(J)}{dt}$	$\frac{df_4(J, N)}{dt}$	$f_{21}(J, L)$
15	$f_{21}(F1, J)$	$\delta(f_3(I, N))$	$\delta(f_3(K, J))$	$\frac{df_3(A1, I)}{dt}$	$\frac{df_3(I, J)}{dt}$	$\delta(f_4(J, M))$
16	$f_{21}(K, L)$	$\frac{df_3(K, L)}{dt}$	$\delta(f_3(K, L))$	$\frac{df_3(K, L)}{dt}$	$\frac{df_3(K, L)}{dt}$	$f_{21}(K, L)$
18	$\delta(f_3(I, J))$	$\frac{df_3(I, J)}{dt}$	$\delta(f_3(I, J))$	$\frac{df_3(I, J)}{dt}$	$\frac{df_3(I, J)}{dt}$	$f_{21}(I, J)$
20	$\delta(f_3(I, J))$	$\frac{df_3(I, J)}{dt}$	$\delta(f_3(I, J))$	$\frac{df_3(I, J)}{dt}$	$\frac{df_3(I, J)}{dt}$	$\frac{df_3(I, J)}{dt}$
22	$\delta(f_3(K, L))$	$\delta(f_3(I, K))$	$\delta(f_3(K, L))$	$\frac{df_3(K, L)}{dt}$	$f_4(B, I)$	$f_{24}(K, L)$
24	$\delta(f_3(I, N))$	$\delta(f_3(D1, L))$	$\delta(f_3(D1, I))$	$\frac{df_3(D1, I)}{dt}$	$\delta(f_3(B1, E1))$	$\delta(f_3(D1, F))$
25	$\delta(f_3(K, L))$	$\delta(f_3(K, L))$	$\delta(f_3(K, L))$	$\frac{df_3(K, L)}{dt}$	$\frac{df_3(K, L)}{dt}$	$\frac{df_3(K, L)}{dt}$
26	$f_{21}(K, L)$	$\frac{df_2(L)}{dt}$	$\delta(f_3(K, L))$	$\frac{df_3(K, L)}{dt}$	$\frac{df_3(K, L)}{dt}$	$f_{21}(K, L)$
27	$\delta(f_3(B, L))$	$\frac{df_3(K, L)}{dt}$	$f_3(K, L)$	$\frac{df_3(K, L)}{dt}$	$\frac{df_3(K, L)}{dt}$	$f_{21}(K, L)$
30	$\delta(f_3(K, M))$	$\delta(f_3(K, M))$	$\delta(f_3(K, M))$	$\frac{df_3(K, L)}{dt}$	$\delta(f_2(E))$	$\delta(f_4(A, D))$
43	$f_{21}(F, H)$	$\delta(f_4(D1, F))$	$\delta(f_3(F, N))$	$\frac{df_4(E1, F)}{dt}$	$\frac{df_4(D1, F)}{dt}$	$f_{21}(F, H)$
45	$\frac{df_4(D1, F)}{dt}$	$\frac{df_4(D, F1)}{dt}$	$f_{23}(E, F1)$	$\frac{df_4(D, F1)}{dt}$	$\frac{df_4(B1, F)}{dt}$	$f_{19}(E1, F1)$
46	$\delta(f_4(F, G1))$	$\frac{df_4(F1, G)}{dt}$	$f_4(D1, F)$	$f_{24}(F, G1)$	$\delta(f_4(A1, L))$	$f_4(E, I)$

Table 7.5: Most important features for distinguishing pair-wise between the four temporal segments neutral, onset apex and offset.

System	Cl. Rate	Recall	Precision	F1-measure
No temporal analysis	0.876	0.772	0.402	0.521
Hybrid SVM-HMM	0.943	0.756	0.653	0.692

Table 7.6: Comparison of the various event detection methods.

the video, scoring an AU event. We applied this approach to the output of the SVM-HMM temporal analysis system, and compare this method with the method without temporal analysis (described in section 6.4.2). The results of this test are shown in table 7.6.

From table 7.6 we can see that the hybrid SVM-HMM with GentleBoost feature selection event detection method outperforms the AU activation event method with a considerable margin. We want to take a closer look at this method for event detection, to be able to compare it with the results presented in 6.4.2. Table 7.7 gives a detailed breakdown of the performance of the SVM-HMM method per AU. In effect, we can only hope to increase the precision of the event detection here, compared to the AU activation adaptive threshold method presented in section 6.4.2, as we do not scrutinise the videos that were predicted to have no active frames and thus might have been false negatives in terms of AU event detection. But, if we take a look at table 6.6, the recall is already quite high and it is exactly the precision that we would like to increase. Indeed, our results show that this approach is extremely effective, increasing ϕ_1 by 17.1% and p_r by 25.1%!

AU	Cl. Rate	Recall	Precision	F1-measure
1	0.971	0.909	0.800	0.851
2	0.947	0.760	0.731	0.745
4	0.909	0.684	0.722	0.703
5	0.959	0.684	0.765	0.722
6	0.938	0.778	0.700	0.737
7	0.959	0.533	0.727	0.615
9	0.951	0.733	0.579	0.647
10	0.938	0.706	0.545	0.615
12	0.922	0.824	0.467	0.596
13	0.963	0.929	0.619	0.743
15	0.959	0.667	0.667	0.667
16	0.918	0.778	0.467	0.583
18	0.934	0.562	0.500	0.529
20	0.967	0.800	0.706	0.750
22	0.955	0.933	0.583	0.718
24	0.955	0.600	0.643	0.621
25	0.942	0.962	0.909	0.935
26	0.831	0.581	0.391	0.468
27	0.967	0.933	0.667	0.778
30	0.951	0.500	0.583	0.538
43	0.963	0.867	0.650	0.743
45	0.938	0.907	0.951	0.928
Avg:	0.943	0.756	0.653	0.692

Table 7.7: Subject independent cross validation results for AU activation event detection after identification of the temporal segments of AUs. Results are for 244 examples taken from the MMI Facial Expression Database

Chapter 8

Emotion recognition

8.1 Emotion recognition: conscious vs. unconscious reasoning approaches

The ability to detect and understand facial expressions and other social signals of someone with whom we are communicating is the core of social and emotional intelligence [Goleman, 1995]. Human Machine Interaction systems capable of sensing stress, inattention and heedfulness and able to adapt and respond to these affective states of users are likely to be perceived as more natural, efficacious and trustworthy [Pantic and Rothkrantz, 2003].

But what exactly is an affective state? Traditionally the terms *affect* and *emotion* have been used synonymously. Following Darwin, discrete emotion theorists proposed the existence of six or more basic emotions that are universally displayed and recognised (see chapter 2). These include emotions such as happiness, anger, sadness, surprise, disgust and fear.

In the previous two chapters we have focused on recognising the signs of a facial expression, in terms of AUs and their temporal segments. While this is an objective and complete description of facial expressions, they are hard for humans to understand or use in communication. When we talk to each other, we rather say that someone was looking sad than that he had activated AU4 and AU15. Ultimately, we will reach a point where we need to interpret the facial expression, now coded in AUs, in some more abstract terms. In this chapter we will investigate the possibilities to map the recognised AUs to the six basic emotions. We have chosen for this description of facial expressions because

Table 8.1: Rules for mapping AUs to emotions, according to the FACS investigators guide. A||B means ‘either A or B’.

Emotion	AUs	Emotion	AUs
Happy	{12}	Fear	{1,2,4}
	{6,12}		{1,2,4,5,20,
Sadness	{1,4}		25 26 27}
	{1,4,11 15}		{1,2,4,5,25 26 27}
	{1,4,15,17}		{1,2,4,5}
	{6,15}		{1,2,5,25 26 27}
	{11,17}		{5,20,25 26 27}
Surprise	{1}		{5,20}
	{1,2,5,26 27}		{20}
	{1,2,5}	Anger	{4,5,7,10,22,23,25 26}
	{1,2,26 27}		{4,5,7,10,23,25 26}
{5,26 27}	{4,5,7,17,23 24}		
Disgust	{9 10,17}		
	{9 10,16,25 26}		{4,5 7}
	{9 10}		{17,24}

there is a fair number of applications that we could implement if we are able to detect emotions (see chapter 2). Also, existing facial expression recognition systems often recognise emotions so this allows a comparison of our work with that of others.

Classic cognitive scientist studies like the EMFACS (emotional FACS), suggest that it is possible to map AUs onto the basic emotion categories using a finite number of rules (as suggested in the FACS investigators guide [Ekman et al., 2002], table 8.1). This effectively suggests that facial expressions are decoded at a conscious level of awareness. It suggests that we first consciously observe exactly what parts of the face move or change appearance and then infer at a conscious level that this means that a certain emotion is shown. Alternative studies, like the one on ‘the thin slices of behaviour’ [Ambady and Rosenthal, 1992], suggest that human expressive nonverbal cues such as facial expressions are neither encoded nor decoded at an intentional, conscious level of awareness. In turn, this finding suggests that biologically inspired classification techniques like artificial neural networks (ANNs) may prove more suitable for tackling the problem of (basic) emotion recognition from AUs as such techniques emulate unconscious human problem solving processes in contrast to rule-based techniques, which are inspired by conscious human problem solving processes.

Recent work on emotion detection using biologically inspired algorithms used ANNs [Fasel et al., 2004], SVMs [Bartlett et al., 2004], Bayesian Networks [Cohen et al., 2003, Zhang and Ji, 2005] and Hidden Markov Models (HMMs) [Cohen et al., 2003]. Recent work on facial AU detection using biologically

inspired algorithms has used similar techniques: ANNs [Tian et al., 2005], SVMs [Bartlett et al., 2004, Valstar and Pantic, 2006], and Bayesian Networks [Zhang and Ji, 2005]. Recent work on AU and emotion detection that used algorithms inspired by human conscious problem solving processes includes rule-based systems [Pantic and Patras, 2006], case-based reasoning [Pantic and Rothkrantz, 2004], and latent semantic analysis [Fasel et al., 2004]. See chapter 2 for an exhaustive overview of AU detection systems.

The goal of the study presented in this chapter is a twofold. First we want to investigate whether a two-step approach to emotion recognition, where both the facial feature extraction and the AU recognition precede the emotion prediction, attains similar recognition rates as a single-step approach in which the recognition of emotions is conducted based directly upon the extracted facial features. Detection of AUs as a first step in facial expression detection has several advantages. AUs are independent from high level interpretations in terms of emotions or moods. They also cause a dimensionality reduction, as all expressions can be described using only 44 attributes, namely the 44 AUs. In contrast to one-step expression detection, AU detectors can be trained independently of the facial expression shown. Hence, in order to train an AU detector the training data does not need to contain examples of all 7000 frequently occurring facial expressions. On the other hand, the main reason to presuppose that a single-step approach could perform better is the error accumulation inherent in multiple-step approaches.

The second goal is to investigate the suggestion implicitly made by recent alternative studies in psychology that biologically inspired classification techniques like ANNs are more suitable for tackling the problem of emotion recognition than logic inspired classifiers such as rule-based systems.

The features used for this study are f_1 to f_{12} (see equations 5.12 to 5.19 in section 5.3). The action unit recognition procedure was analogous to that described in section 6.3. See section 6.4.1 for the performance of this AU detector on various databases. Figure 8.1 shows an outline of the systems we will compare in this chapter.

8.2 One-step emotion recognition

We decided to keep our one-step approach to emotion recognition analogous to the method used for AU detection. Thus, we use the features f_1 to f_{12} (see equations 5.12 to 5.19 in section 5.3) and used

gentleSVMs to recognise the emotions. To solve our six emotion detection problem we used a one-versus-one multi-class SVM classifier. Feature selection by GentleBoost turned out to be particularly efficient. On average, 7 of a total of 1260 features were picked as the most informative features in one-vs-one emotion classification. It also resulted in a higher emotion recognition result. The average F1-measure increased by 17.0% (see Table 8.2).

8.3 Two-step emotion recognition

In the two-step approach to emotion recognition from face image sequences, both the facial feature extraction and the AU recognition precede the emotion prediction (see Fig. 8.1). We use the AU detector described in chapters 6 and 7, that can detect 23 AUs. The choice of 23 AU categories was determined by the AUs that can be encoded based upon the utilised features defined in section 5.3. As we can see from table 8.1, the set of AUs that we are able to detect contains all but two of the components of expressions (i.e., micro-events) that seem to be hardwired to emotions [Ekman et al., 2002].

To perform emotion classification based on AU predictions (the second stage of our two-step emotion recognition approach), we experimented with both the biology and the logic inspired classification engines. The logic inspired recognition engine is a rule-based system that maps the 15 AUs onto the 7 emotion categories. The utilised rules are the EMFACS rules suggested by Ekman and colleagues in the FACS investigators guide (see Table 8.1, [Ekman et al., 2002]). The biology inspired recognition engine that we have experimented with was an ANN with 3 hidden layers, each of which had 23 nodes, and one output layer containing 6 nodes, one for every emotion. All neurons used the log sigmoid evaluation function. The number of layers, neurons, and the evaluation function were empirically determined.

8.4 Experiments

For this study we have used data from the Cohn-Kanade database. The database consists of a total of 487 recordings of 97 subjects. From this set, we selected 156 image sequences of 66 subjects. Image sequences were included in this validation set if two experts decided by consensus on the basic emotion displayed in the samples in question.

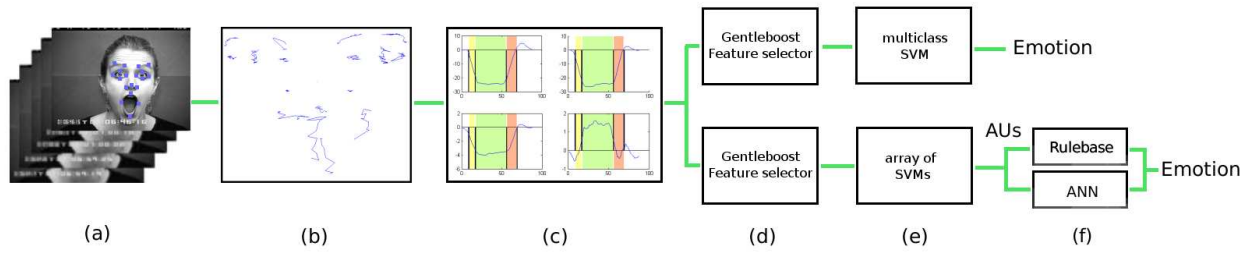


Figure 8.1: Overview of the automatic AU and emotion detection system: (a) input image sequence with tracked facial points, (b) tracking results, (c) the four most important features for recognition of AU2 shown over time, (d) GentleBoost is used to select the most important features, (e) which are subsequently fed to (mc-)SVMs, (f) for two step approach: emotion detection from AUs by Artificial Neural Networks or a rulebase.

Table 8.2: Results of one-step emotion recognition, *clr* is classification rate, *rec* is recall and *pr* is precision: A) classification of features to emotions by a multi-class SVM, B) one-step emotion classification without feature selection.

Emotion	A				B			
	clr	rec	pr	ϕ_1	clr	rec	pr	ϕ_1
Anger	0.948	0.643	0.750	0.692	0.923	0.549	0.608	0.577
Disgust	0.948	0.920	0.793	0.852	0.926	0.598	0.652	0.624
Fear	0.922	0.760	0.760	0.760	0.901	0.495	0.662	0.566
Happy	0.967	0.972	0.897	0.933	0.907	0.757	0.822	0.788
Sadness	0.935	0.619	0.867	0.722	0.903	0.307	0.613	0.409
Surprise	0.967	0.938	0.909	0.923	0.966	0.829	0.938	0.880
Average	0.948	0.809	0.829	0.819	0.903	0.620	0.681	0.649

Table 8.3: Results of two-step emotion recognition, *clr* is classification rate, *rec* is recall and *pr* is precision: A) classification of manually labelled AUs to emotions by rules, B) automatically detected AUs classified to emotions by rules, C) Neural Networks classifying automatically detected AUs into emotions .

Emotion	A				B				C			
	clr	rec	pr	ϕ_1	clr	rec	pr	ϕ_1	clr	rec	pr	ϕ_1
Anger	0.967	0.714	0.909	0.800	0.889	0.214	0.333	0.261	0.915	0.500	0.539	0.519
Disgust	0.980	0.960	0.923	0.941	0.856	0.920	0.535	0.677	0.935	0.760	0.826	0.792
Fear	0.980	0.960	0.923	0.941	0.889	0.680	0.654	0.667	0.895	0.720	0.667	0.693
Happy	1.00	1.00	1.00	1.00	0.941	0.917	0.846	0.880	0.948	0.917	0.868	0.892
Sadness	0.980	0.952	0.909	0.930	0.882	0.191	0.800	0.308	0.889	0.571	0.600	0.585
Surprise	1.00	1.00	1.00	1.00	0.941	0.844	0.871	0.857	0.974	0.938	0.938	0.938
Average	0.985	0.931	0.944	0.938	0.900	0.626	0.673	0.649	0.926	0.734	0.740	0.737

The first goal of this study is to investigate the performance of a two-step approach in which AUs are detected automatically in the first step and the complex expressions, in this case six basic emotions, in the second step. We compare this two-step approach with the one-step emotion detection where we feed the features directly into a multiclass SVM to detect emotions. Both two-step approaches use binary SVMs to detect AUs first. In the second step we use either an ANN or a rulebase to map the AUs to emotions. This second step in the two-step approach also gives us the data needed for the second goal of our study: evaluating whether the classic or alternative psychological studies made a correct assumption, i.e., whether biology inspired techniques indeed outperform logic inspired ones. Figure 8.1 shows a diagram of the discussed systems.

All experiments were performed using a leave-one-subject-out cross validation strategy. The parameters for the trained (Gentle-)SVMs were obtained as described in section 6.2. Results for one-step emotion detection are shown in table 8.2.

The ANN used to detect emotions from AUs in the two-step approach was evaluated using a leave-one-subject-out cross validation as well. The rulebase used to detect emotions from AUs in the two-step approach was directly applied to the output of the automatic AU detector as the rules are fixed and do not need any training. Table 8.3 show the two-step approach results for the logic inspired rulebase and the biologically inspired ANN, respectively. To test the “goodness” of our rulebase, we also applied the rulebase on manually coded AUs. As we can see from table 8.3, for the two-step approach the ANN performs best. The F1-measure obtained by ANNs is 8.8% higher than that obtained by applying the rule-base to the automatically detected AUs.

Table 8.3 part B shows that results for the logic inspired approach deteriorate significantly when the rules are applied on automatically detected AUs instead of manually labelled AUs. Obviously, the rulebase is a very rigid, nonadaptive system that is sensitive to noise in the input. The fact that the rule based classifier cannot learn and has no means to compensate for known weaknesses in the AU detector (such as the poor classification of AU15 which influences sadness recall, see table 6.4) contributes to a performance decrease. This is not the case for the biologically inspired ANN, as part C of table 8.3 clearly shows. This suggests that the alternative studies in psychology offer a model for high-level facial behaviour interpretation that is more suitable for computer-based analysis than the model offered by classic studies.

When we compare tables 8.2 and 8.3, we observe that the two-step approach performs worse than the

one-step approach. However, the difference between performing two-step emotion recognition with ANNs and performing one-step emotion recognition is not that big (an F1-measure difference of 8.2%). The interpretation free description of expressions in terms of AUs, the decreased dimensionality of the classification problem, and the ability to analyse the temporal segments of the AU components that make up the emotive expression could be considered more valuable than the increase in accuracy.

These results also clearly confirm the validity of our claim that it is possible to use AUs to automatically recognise more complex expressions. Although not unexpected, this is an important result as without it there would be little point in detecting AUs in the first place.

8.5 Application: Emotionally Aware Painting Fool

As an application to emotion detection, we combined the two-step emotion detection system described in this chapter with The Painting Fool, created by Dr. Simon Colton. The Painting Fool is capable of painting someone's portrait given an input image. The idea was to make the Painting Fool appreciate its subject more by using information about the emotion shown by the subject. To do so, we first recorded a small video of a subject showing one of the six basic emotions using a webcam. We then proceeded with analysing the activated AUs using the techniques proposed in chapters 6 and 7. Based on the AU activation detection and the ANN described in section 8.3, we decided what emotion was shown. For maximal visual effect we wanted to work with an image in which the emotion was shown very strong, therefore we identified the frame in which the most AUs were in their apex phase. We called this frame the *apex frame*. In order to allow the Painting Fool to enhance the image around certain facial features, such as the eyes, the nose, and the mouth, we also passed the positions of the tracked points in the apex frame to it. We coined the combination of our facial expression analysis system and The Painting Fool: *The Emotionally Aware Painting Fool*¹.

The Painting Fool bases the portrait on the image provided from the emotional modeling software, and chooses its art materials, colour palette and abstraction level according to the emotion being expressed. For instance, if it was told that the person was expressing happiness, it chose vibrant colours, and painted in simulated acrylic paints in a slapdash way. If, on the other hand, it was told that the person was sad, it chose to paint with pastels in muted colours. To make the painting look

¹The Emotionally Aware Painting Fool, http://www.thepaintingfool.com/projects/mi_competition/index.html, date of access: February 29 2008



Figure 8.2: Apex frame captured of the author showing his feelings during the Machine Intelligence Competition 2007.

more like the sitter, the areas around the eyes, nose and mouth were painted in greater detail. This was possible as the facial expression analysis system passed the facial point tracking information on to The Painting Fool.

The entire system was integrated on a laptop and demonstrated during The Machine Intelligence Competition 2007. This is an annual event sponsored by Electrolux and organised by the British Computer Society, as part of their SGAI International Conference on Artificial Intelligence. The competition awards a prize to the best live demonstration of Artificial Intelligence software which shows the most progress towards machine intelligence. On December 11th 2007, the competition was held at Peterhouse College at the University of Cambridge. We entered our “Emotionally Enhanced Painting Fool” system, and we were lucky enough to win the competition. The team consisted of Dr. Simon Colton, Ir. Michel Valstar and Dr. Maja Pantic. Unfortunately Dr. Pantic was ill on the competition day and could not attend. Figure 8.3 shows the portrait painted of this thesis’ author during the demonstration, where he showed the emotion ‘disgust’. The original is shown in Figure 8.2.

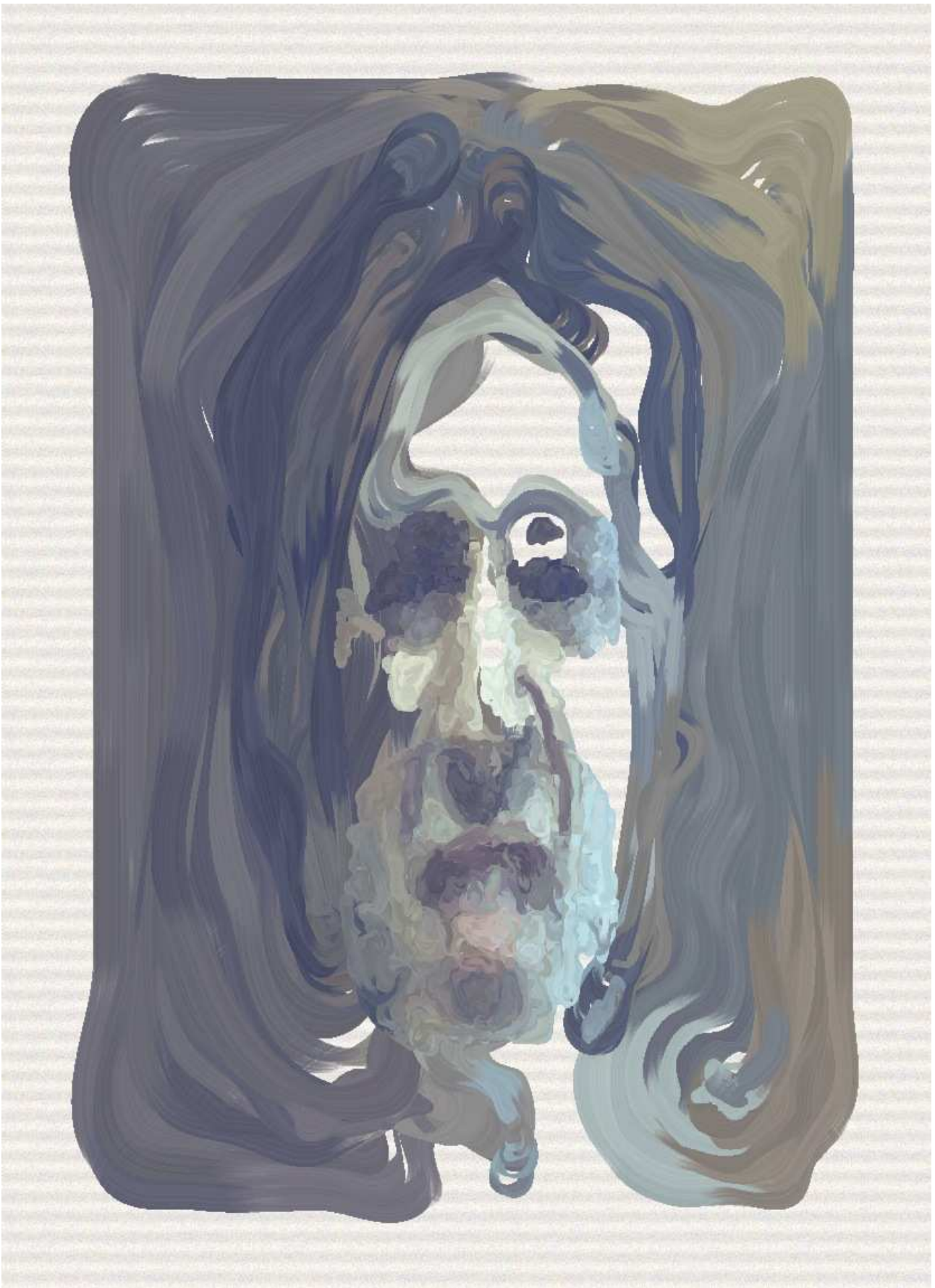


Figure 8.3: Portrait of the author showing his feelings during the Machine Intelligence Competition 2007.

Chapter 9

Applications to human facial behaviour understanding

”No mortal can keep a secret. If his lips are silent, he chatters with his fingertips; betrayal oozes out of him at every pore.”

Sigmund Freud

The detection and temporal analysis of atomic facial actions (i.e., AUs), has many applications to behaviour analysis. Their relevance has already been studied intensively by cognitive scientists [Ambadar et al., 2005, Bassili, 1978, de C. Williams, 2006, Ekman and Rosenberg, 2005, Hess and Kleck, 1990, Russell and Fernandez-Dols, 1997]. However, few works exist that use automatic analysis of facial expressions to analyse human behaviour analysis.

In section 9.1 we will discuss previous work related to understanding of human spontaneous behaviour using FACS. While the majority of the efforts in automatic facial expression recognition have been designed to deal with unnatural, posed expressions, the works discussed in section 9.1 have focused on the understanding of facial expressions that occur involuntarily, usually in real-world situations. In the literature, these expressions are referred to as *spontaneous* expressions, while the expressions that are performed deliberately are referred to as *posed* expressions. In this thesis we will adhere to that terminology.

Furthermore, in this chapter we will discuss two studies that we carried out to investigate how well our system for the automatic analysis of AUs and their temporal dynamics can be applied to address

various aspects of behaviour analysis. To wit, in section 9.2 we use the temporal dynamic aspects of AUs and motion patterns of tracked facial points to distinguish posed from spontaneous brow actions and in section 9.3 we investigate the possibilities of a multi-cue approach to posed vs. spontaneous smile detection.

9.1 Related work in automatic spontaneous facial expression analysis

In the same period in which the work presented in this thesis was carried out, there have been only two other research groups which addressed the recognition of spontaneous facial expressions. The first group to address this issue was that of Bartlett from the University of California, San Diego in 2003. This was followed shortly after by the group of Jeffrey F. Cohn at the University of Pittsburgh/Carnegie Mellon University who published their first paper on this topic in 2004.

The work by Bartlett presented in 2003 [Bartlett et al., 2003] followed a similar approach as presented by that group for the detection of posed expressions [Bartlett et al., 1999]. They mentioned that one of the greatest challenges in spontaneous facial expression analysis is being able to cope with 3D rigid head motion, especially out of plane rotation. To this end they used a 3D face model (of which unfortunately no details or references were given), to separate rigid from non rigid facial motion and register the face to a frontal view. However, 8 facial points had to be manually annotated in every frame to estimate the head pose. The back-projected frontal view of the face, retrieved after face registration, was convolved next with a bank of Gabor filters. The magnitude responses of the filters were used as the input to first train SVMs. The output of the SVMs (the distance to the hyperplane for a given example) was then input to a HMM in order to incorporate a model of time in the AU recognition system. The proposed system was able to recognise AU45 (blink), AU1+AU2 (brow raiser) and AU4 (brow lowerer), with 98.2%, 75%, and 78.2% classification rate, respectively. They did not report on recall, precision or F-measure. The utilised dataset was rather unbalanced, so these results are not as good as they seem at first glance. For example, there were 62 brow raise examples of in total 244 examples, which would result in a classification rate of 74.6% if the classifier would not return a single brow raise. This means that the reported 78.2% is very low indeed.

The paper presented by Cohn et al. in 2004, [Cohn et al., 2004b], was based on the automated facial image analysis system (AFA) presented earlier by the same group in [Tian et al., 2001], which was described in section 2.2.1 of this thesis. The system was extended with a headtracker that uses a cylindrical semi-3D face model to estimate the head pose and register the face accordingly

[Xiao et al., 2003]. This is essential for dealing with spontaneous expressions, where a considerable amount of rigid head motion is to be expected. The head pose values were combined together with the geometry based features and the Gabor filter appearance based features into one feature vector. Discriminant analysis was used as the classification method. The system was able to recognise two different AU combinations; AU4 (brow lowerer) and AU1+AU2 (brow raiser) with 60% and 82% classification rate, respectively, on a dataset taken from 101 different subjects. Although the eyebrows are the facial features which are easiest to track and the AU combinations AU1+AU2 and AU4 represent one of the easiest brow actions to discriminate from each other, the results remain significant considering that this was only the second attempt to recognise spontaneous facial actions ever.

The system proposed by Bartlett et al. in 2006 [Bartlett et al., 2006] was both an improvement and a simplification of their 2003 version. The features and the SVM classification remain the same, but they are now able to detect 20 AUs, with high accuracy. However, the system presented in this work does not perform any 3D tracking and is therefore not able to deal with out-of-image-plane head rotations. Translations and in-plane rotations are dealt with by detecting the irises and registering the face image accordingly. Also, modelling of time by a HMM was removed.

In 2007 Lucey et al. presented a system that was able to detect 4 AU combinations in spontaneous expressions that contained large head rotations, including severe out-of-image-plane rotations [Lucey et al., 2007]. This system employed Active Appearance Models (AAMs), which track facial features using shape models and appearance based features simultaneously. Inspired by the hybrid tracking approach, the authors use geometric features (the positions of the points in the tracked face mesh) and appearance based features (pixel intensities of locations around facial points, both with and without shape information) to recognise 4 different AU combinations in the upper face in spontaneous data. An important shortcoming of AAMs is that they have to use a separate shape model for every subject that the system will be used on. This means that it cannot directly operate on a novel subject; it first needs to be given a face model of the new subject. Their proposed system could distinguish between the classes AU1, AU1+AU2, AU4 and AU5, with an average classification rate of 70.5% per class.

Cohn's group used the same system to detect expressions of pain [Ashraf et al., 2007]. The data consisted of video clips for 129 subjects who suffered from shoulder pains. When visiting a physiotherapist they had to rotate their shoulder and report their pain on a Visual Analog Scale. The videos were also rated by expert pain identifiers and scored on a 5-point Likert scale by them. The proposed system did not recognise AUs before detecting pain. Instead, it used the same features as presented in [Lucey et al., 2007] directly as input to a linear SVM to detect the pain. The authors reported a

81.2% hit rate and a 81.2% classification rate when trying to score the same as the expert.

The work proposed by Littlewort et al. [Littlewort et al., 2007] on the other hand, does use the AU detection to analyse pain. The goal of their system was to distinguish expressions of posed pain from expressions of real pain. To achieve this, they use the same system as described in [Bartlett et al., 2006]. That system used a separate SVM for each AU to be detected. To distinguish between posed and fake expressions of pain, they use the real-valued output of the SVMs. These outputs were fed to another, higher-level SVM, which decided if the expression of pain was real. The system was reported to attain a level of 72% accuracy, compared with 52% accuracy by naive human judges.

9.2 Posed vs. spontaneous brow actions: automatic detection

This section reports on a method that we propose for automatic discrimination between posed and spontaneous facial expressions using temporal dynamics of brow actions. We focus on brow actions in this section because they are frequent in the repertoire of human facial behaviour, they are relatively easy to track even under challenging conditions and we have a large number of data examples of both spontaneous and posed brow actions.

The brows are often lowered in displays of affective states like anger and fear and are often raised in displays of surprise [Ekman et al., 2002]. Brow lowering is also frequently present in displays of psychological and cognitive states like pain [de C. Williams, 2002], fatigue [Veldhuizen et al., 2003], concentration and puzzlement [Cunningham et al., 2004]. Brow raising is typical for various social signals including greetings [Kendon, 1973], interest [Flecha-Garcia, 2001], and (dis-)agreement [Cunningham et al., 2004]. It may also provide emphasis for speech acts, or contribute to the regulation of turn-taking (like in query) in social interactions [Flecha-Garcia, 2001].

To capture brow action dynamics, we track 8 characteristic facial points in frontal face video and compute their displacements. The points in question are $B, B1, D, D1, E, E1, H$ and $H1$ (see Fig. 4.1): the inner and the outer corners of the eyebrows ($D, D1, E, E1$), the inner corners of the eyes ($B, B1$), and the outer corners of the nostrils ($H, H1$). Due to data privacy constraints, we had to use two different trackers to track the facial points in the videos. The first, used for the confidential data by prof. Jeffrey Cohn's group in Pittsburgh, is the FACE-III tracker created by Xiao et al. [Xiao et al., 2003]. The second tracker is the Particle Filtering with Factorised Likelihoods tracker presented in chapter 5, which was used to track the facial points in all non-confidential videos.

Using the tracking data, we first detect the presence (i.e., activation) of AU1, AU2 and AU4. For each

activated AU, we determine the temporal segments (neutral, onset, apex, and offset). To detect the activated AUs and their temporal segments, we use the AU detector and temporal segment analyser described in chapters 6 and 7, which combines GentleBoost ensemble learning, SVMs, and HMMs.

We compute further a set of mid-level feature parameters for every temporal segment of each activated AU. These include the segment duration, the mean and the maximum displacements of the four brow points in x - and y -directions, the maximum velocity in x - and y -directions, and the asymmetry in the displacement of the brow points. We also compute the 2nd order polynomial functional representation of the displacements of the brow points and the order in which the AUs have been displayed.

We use GentleBoost to learn the most informative parameters for distinguishing between spontaneous and posed AUs and use these to train a separate Relevance Vector Machine (RVM) [Tipping, 2000] for each temporal segment of each of the three brow-action-related AUs (i.e. 9 GentleRVMs in total). The outcomes of these 9 GentleRVMs are then combined and a probabilistic decision function determines the class (spontaneous or posed) for the entire brow action.

When trained and tested on a set containing 60 examples of posed facial displays from the MMI Facial Expression database [Pantic et al., 2005b], 59 examples of posed facial displays from the Cohn-Kanade Facial Expression database [Kanade et al., 2000], and 70 examples of spontaneous facial displays from the DS118 dataset [Rosenberg et al., 1998], the proposed method attained a 90.7% correct recognition rate when determining the class (spontaneous or posed) of an input facial expression.

In the following sections we will give a detailed explanation of our posed vs. spontaneous brow action recognition system. Section 9.2.1 describes how we extract a set of geometric features. In section 9.2.2 we will then describe our classification strategy, that uses an array of Relevance Vector Machines to attain a probabilistic posterior probability of the classes posed and spontaneous brow action. Finally, in section 9.2.3 we provide a detailed evaluation of our proposed method.

9.2.1 Mid-level feature parameters

Our choice of mid-level feature parameters to be used for automatic discrimination between spontaneous and posed brow actions is largely influenced by a number of studies in psychology on spontaneous (produced in a reflex-like manner) and volitional (deliberately produced) smiles. We believe that these parameters could very well be characteristic for posed or spontaneous brow actions too.

Many of these features describe the temporal dynamics of facial actions. The body of research in cognitive sciences, which suggests that the temporal dynamics of human facial behaviour (e.g.

the timing and duration of facial actions) are a critical factor for interpretation of the observed behaviour, is large and growing [Ambadar et al., 2005, Bassili, 1978, Hess and Kleck, 1990]. Facial expression temporal dynamics are essential for categorisation of complex psychological states like various types of pain and mood [de C. Williams, 2002]. They are also the key parameter in differentiation between posed and spontaneous facial expressions [Ekman, 2003, Ekman and Rosenberg, 2005, Hess and Kleck, 1990]. For instance, it has been shown that spontaneous smiles, in contrast to posed smiles (such as a polite smile), are slow in onset, can have multiple AU12 apices (multiple peaks in the mouth corner movement), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1s [Cohn and Schmidt, 2004]. Below we summarise the most important findings:

Intensity: The early study of Ekman and Friesen on felt and false smiles suggested that the intensity of contractions of the zygomaticus major (i.e., the muscle running under the cheek that is responsible for lip corner retraction in smiles) is important in distinguishing between felt and ‘phony’ smiles [Ekman and Friesen, 1982]. Cohn and Schmidt further confirmed this finding, reporting that posed smiles have greater amplitude than spontaneous smiles [Cohn and Schmidt, 2004, Schmidt et al., 2006].

Duration: Ekman and Friesen found that spontaneous smiles usually last between $\frac{2}{3}$ and 4 seconds [Ekman and Friesen, 1982]. Hess and Kleck extended these findings and found that deliberate smiles are shorter in total duration than spontaneous smiles [Hess and Kleck, 1990]. Cohn and Schmidt further confirmed these findings [Cohn and Schmidt, 2004, Schmidt et al., 2006]. Fogel et al. [Fogel et al., 2006] found that even a complex family of different smiles can be defined based on the differences in duration and amplitude of the relevant smiles.

Trajectory: Ekman and Friesen reported that in deliberate expressions the onset is often abrupt, the apex held relatively long, and the offset is either abrupt and/or appears irregular rather than smooth [Ekman and Friesen, 1982]. Hess and Kleck confirmed these observations and reported that in comparison to deliberate smiles, spontaneous smiles are slower in onset and offset time [Hess and Kleck, 1990]. Cohn and Schmidt reported that spontaneous smiles are slow in onset [Cohn and Schmidt, 2004, Schmidt et al., 2006] and they also found that they may have multiple apices [Cohn and Schmidt, 2004].

Symmetry: Ekman et al. [Ekman et al., 1981] reported that spontaneous smiles are more symmetrical than those made deliberately. In one study they found that smiles in response to watching an amusing film were in 96% of cases symmetrical. This finding has been later confirmed by different researchers [Cohn and Schmidt, 2004, Schmidt et al., 2006].

Co-occurrences: Evidence that in spontaneous smiles the activity of the zygomaticus major is accompanied by the activity of the orbicularis oculi (i.e., the muscle orbiting the eye which when contracted produces crow-feet wrinkles around the outer corner of the eye) dates from the 19th century Duchenne de Boulogne [de Bologne, 1862], in whose honour the smiles incorporating orbicularis oculi activation are called Duchenne’s smiles. Duchenne’s proposal has been revisited many times by researchers like Charles Darwin, Paul Ekman, and many others. Recent studies by Cohn and Schmidt extend these findings and suggest that the activity of zygomaticus major (AU12) may be accompanied not only by the activity of orbicularis oculi (AU6) but also by activity of other AUs that appear either simultaneously with AU12 or follow AU12 within 1s [Cohn and Schmidt, 2004].

As suggested by Ekman [Ekman, 2003], the findings about spontaneous and posed smiles may be extendable to a wider set of facial actions. In this study we present our research to check Ekman’s proposal for the case of brow actions. To the best of our knowledge, this is the first study that checks Ekman’s proposal, as far as brow actions are concerned.

We use our AU activation and temporal analysis system presented in chapters 6 and 7 to detect the temporal phases of the brow AUs AU1, AU2 and AU4. For every temporal segment (neutral, onset, apex, offset) of each activated AU, we calculate a number of mid-level feature parameters based on the displacements of facial fiducial points P_i , where $i = [1 \dots 4]$. For further processing we are interested only in brow actions. Therefore we only consider the displacements of points $D, D1, E, E1$ (Fig. 4.1). Let us name these points P_2, P_3, P_1 and P_4 for convenience (note the order of the points). For a temporal segment d consisting of n frames, we will have signals $s_{i,t} = \{f_1(P_i, t), f_2(P_i, t)\}$ (see the feature definitions in section 5.3), where $t = [1 \dots n]$ is a frame of the temporal segment d . The function $f_1(P_i, t)$ returns the x -position of point P_i at time t and $f_2(P_i, t)$ the y -position of that point at that time. Thus, for the entire temporal segment d , we will have a set of signals $S_i = \{s_{i,1}, \dots, s_{i,n}\}$ for each facial point $P_i, i = [1 \dots 4]$, resulting in a total of $N_d = 4 \times n$ signals.

For each signal S_i corresponding to point tracking result P_i , and for both the y - and x -directions, we first compute the maximum and the mean point displacement (relating to the *intensity* of brow actions), and the maximum and the mean velocity of these displacements (relating to the *speed* and the *trajectory* of AU activations). We do so as follows:

$$f_{m1}(S_i) = \max(S_i) \quad (9.1)$$

$$f_{m2}(S_i) = \frac{\sum_i S_i}{n} \quad (9.2)$$

$$f_{m3}(S_i) = \max\left(\frac{s_{i,t} - s_{i,t-1}}{v}\right) \quad \forall t \in [1 \dots n] \quad (9.3)$$

$$f_{m4}(S_i) = \frac{\sum_t (s_{i,t} - s_{i,t-1})}{(n-1)v} \quad (9.4)$$

where v is again the framerate of the source video. Next, for the pairs of the brow points P_1, P_4 and P_2, P_3 , we compute a measure of symmetry. We say that a given brow action is symmetric in x -direction if the relevant brow points move an equal distance towards each other or away from each other. Otherwise, the brow action in question is asymmetric. We say that a given brow action is symmetric in y -direction if the relevant brow points traverse an equal distance either upwards or downwards. Otherwise, the brow action in question is asymmetric. As the measure of symmetry in y -direction we use mid-level feature parameter f_{m5} defined by equation (9.5) and as the measure of symmetry in x -direction we use mid-level feature parameter f_{m6} defined by equation (9.6),

$$f_{m5}(P_i, P_j) = \sum_t \{|f_1(P_i, t) - f_1(P_j, t)|\} \quad (9.5)$$

$$f_{m6}(P_i, P_j) = \sum_t \{|f_2(P_i, t) + f_2(P_j, t)|\} \quad (9.6)$$

Next, we describe the overall displacement of a facial fiducial point within the temporal segment d as a function of time $g(t) = at^2 + bt + c$, where $t = [1 \dots n]$ is a frame of the temporal segment d and $g(t) = s_{i,t}$, analogous to the Local-Time-Parameterised features described in section 5.3.3. Thus, for each S_i , and for both the y - and x -directions, we define the following mid-level feature parameters (relating to the *trajectory* of AU activations):

$$f_{m7}(S_i) = a \quad (9.7)$$

$$f_{m8}(S_i) = b \quad (9.8)$$

$$f_{m9}(S_i) = c \quad (9.9)$$

Finally, the total duration of a temporal segment d (i.e., n) and the occurrence order o of the temporal segment within the entire image sequence ($o = 1$ if d was the first temporal segment of the first activated AU, otherwise $o > 1$) are used as mid-level feature parameters as well.

$$f_{m10}(d) = n \quad (9.10)$$

$$f_{m11}(d) = o \quad (9.11)$$

Thus, for each temporal segment of an activated AU, we calculate in total 62 features. Remember that we use only four points, and that we compute the features $f_{m7} \dots f_{m9}$ for both the x and y components of the trackings.

9.2.2 Classification strategy

The tracking schemes that we utilise to track facial characteristic points in an input face image sequence include a particle-filtering-based and an optical-flow-based tracking algorithm. Although we have aimed to achieve ‘best effort’ results, meaning that some of the input face image sequences have been tracked more than once using slightly different initialisation parameters to remove any errors that the tracker has made in earlier trials, noisy output from the tracking algorithms should still be expected. In the case of the PFFL point tracker the noise in the output (occurring due to the particle filtering nature of the tracker) is filtered out by applying a temporal filter that removes spurious peaks. However, since the smoothed output of the tracker for a frame at time t is computed by taking the mean of the tracker outputs in the range $[t-2, t+2]$, some noise remains present in the output. For the FACE-III head tracker, the noise in the facial point tracking data occurs mostly due to inaccuracies in registration as well as due to the applied re-registration process.

This noise in point tracking data is propagated further, resulting in noisy feature-based and mid-level parametric representations of the input data. Eventually, this will influence the predictions of whether a given temporal segment belongs to a spontaneous or to a deliberate brow action. In order to deal with the imperfect data and generate predictions about whether a given temporal segment belongs to a certain class (spontaneous or posed) so that the confidence measure associated with it varies in accordance with the accuracy of the input data, we employ Relevance Vector Machines (RVMs).

A RVM classifier is a probabilistic sparse kernel model identical in functional form to a SVM classifier [Tipping, 2000]. In their simplest form, RVMs attempt to find a separating hyperplane defined as a

weighted combination of a few relevance vectors that divides data samples of two different classes. In RVM, a Bayesian approach is adopted for learning, where a prior is introduced over the model weights, governed by a set of hyperparameters, one for each weight. The most probable values of these hyperparameters are iteratively estimated from the data. The solution is sparse because the posterior distributions of many of the weights are sharply peaked around 0.

Unlike the SVM, the nonzero weights of RVM are not associated with examples close to the decision boundary, but rather appear to represent prototypical examples of classes. These examples are called relevance vectors and, in our case, they can be thought of as representative displays of either posed or spontaneous brow actions.

The main advantage of an RVM is that while its generalisation capacity is comparable to that of an equivalent SVM, it uses substantially fewer kernel functions. Furthermore, predictions in RVM are probabilistic, in contrast to the deterministic decisions provided by SVM. Therefore it is more natural to use RVMs in a larger probabilistic framework. The major disadvantage of RVMs on the other hand is that they take a very long time to train.

In their original form, RVMs are suitable for solving two-class classification problems. Since for each temporal segment of a brow action we want to determine whether it has been displayed spontaneously or deliberately, our problem is a set of L two-class classification problems. Hence, we use L RVMs, each of which predicts whether a given temporal segment of a specific brow action has been displayed spontaneously or not. In our case $L = 9$ since we have three AUs related to brow actions (i.e., AU1, AU2, AU4), each of which can be in one of the three temporal phases (i.e., onset, apex, offset). We use GentleBoost to select a set of mid-level feature parameters from equations (9.1)-(9.11) that are most informative for distinction between the 9 classes. Then, we train the RVMs to perform binary decision tasks using one-versus-all partitioning of data resulting from the feature selection stage. As a kernel, we use the standard Gaussian radial basis function. For each fold of the cross validation test procedure, the kernel width has been optimised independently of the test data.

Thus, for an input image sequence I in which m temporal segments d_t of brow actions have been identified by the AU temporal analysis system defined in section 7, we will have m predictions c_t (one for each d_t), each of which is associated with a confidence measure $p_t \geq 0$, where $c_t \in \{-1, 1\}$ (i.e., -1 for posed and 1 for spontaneous) and $t = [1 \dots m]$. Then, the class $c \in \{-1, 1\}$ for the entire brow action shown in the input video I is predicted with a confidence measure $P_c \geq 0$ in the following way:

$$\kappa = \sum_{t=1}^m p_t c_t \quad (9.12)$$

$$C(\kappa) = \text{sign}(\kappa) \quad (9.13)$$

$$P(\kappa) = |\kappa| \quad (9.14)$$

9.2.3 Evaluation

In our study, we used a set containing 60 examples of posed facial displays from the MMI Facial Expression database (section 3.4, [Pantic et al., 2005b]), 63 examples of posed facial displays from the Cohn-Kanade Facial Expression database [Kanade et al., 2000], and 139 examples of spontaneous facial displays from the DS118 dataset [Rosenberg et al., 1998]. When we selected this data, we took care that samples of various brow actions were picked from the three databases with the same frequency.

The DS118 dataset has been collected to study facial expression in patients with heart disease. Subjects were 85 men and women with a history of transient myocardial ischemia who were interviewed on two occasions at a 4-month interval. They averaged 59 years of age (std = 8.24) and were predominantly Caucasian. Spontaneous facial expressions were video-recorded during a clinical interview that elicited AUs related to disgust, contempt, and other negative emotions as well as smiles. The brow actions displayed in the data are often very subtle. Due to confidentiality issues, this FACS-coded dataset is not publicly available.

To evaluate the proposed method for automatic discrimination between spontaneous and deliberate brow actions, we used 119 examples of posed brow actions and 70 examples of spontaneous brow actions for which our system for automatic recognition of AUs and their temporal segments generated correct AU event detection results (compared to the ground truth).

Although the aim of this study is not to evaluate the performance of a fully automated system for spontaneous and posed brow action detection, but to investigate whether posed brow actions can be automatically distinguished from spontaneous brow actions based on the temporal dynamics of these actions, we would like to make few remarks about the AU recognition results attained by our AU detector. The version of the AU detector that we used in this study has been trained using only samples of posed facial displays from the MMI database. Hence, it is not surprising that for deliberate facial displays, it achieved high recognition rates. More specifically, from 123 initially used samples of posed brow actions, 119 have been correctly AU-coded by our system, resulting in 96.7% correct

Class	#	Mid-level parameters		
AU1-on	7	$f_{m3}(S_{3,y})$	$f_{m7}(S_{4,x})$	$f_{m1}(S_{1,x})$
AU1-ap	6	$f_{m1}(S_{1,x})$	$f_{m8}(S_{3,y})$	$f_{m3}(S_{2,y})$
AU1-off	11	$f_{m3}(S_{3,y})$	$f_{m11}(d)$	$f_{m1}(S_{1,x})$
AU2-on	3	$f_{m1}(S_{1,x})$	$f_{m2}(S_{4,x})$	$f_{m9}(S_{3,x})$
AU2-ap	2	$f_{m2}(S_{4,x})$	$f_{m11}(d)$	
AU2-off	13	$f_{m1}(S_{1,x})$	$f_{m3}(S_{3,y})$	$f_{m9}(S_{3,x})$
AU4-on	1	$f_{m1}(S_{3,y})$		
AU4-ap	14	$f_{m1}(S_{1,x})$	$f_{m2}(S_{3,y})$	$f_{m11}(d)$
AU4-off	14	$f_{m1}(S_{3,y})$	$f_{m11}(d)$	$f_{m1}(S_{4,x})$

Table 9.1: Mid-level feature parameters that GentleBoost selected as the most informative for determining whether a detected temporal segment d of an activated AU has been displayed spontaneously or not. The 1st column lists the 9 relevant classes, the 2nd column lists the total number of mid-level parameters selected by GentleBoost for the relevant class, and the 3rd column lists the three most informative of the selected parameters defined by equations (9.1)-(9.11)

classification rate.

Second, it is also not surprising that for spontaneous facial actions, an AU detector trained on deliberate facial actions achieves lower recognition rates. From 139 initially used samples of spontaneous brow actions, 70 have been correctly AU-coded by our system, resulting in 50.4% correct recognition rate. Although this is not a very good result, it is promising, especially when one takes into account that the system was not trained on samples of spontaneous brow actions and that current studies on automatic AU-coding of spontaneous facial data reported correct recognition rates ranging from 26% [Bartlett et al., 2006] to 76% [Cohn et al., 2004b] for brow actions.

To evaluate the proposed method for automatic discrimination between spontaneous and deliberate brow actions using the 189 data samples as explained above, we performed a leave-one-subject-out cross validation. For each of the 9 two-class classification problems (i.e., whether onset, apex, and offset of AU1, AU2, and AU4, is spontaneously displayed or not), Table 9.1 lists the mid-level feature parameters that GentleBoost selected as the three most informative for distinction between the 9 classes. As can be seen from Table 9.1, Ekman’s proposal that the findings about posed and spontaneous smiles may be extendable to other facial actions [Ekman, 2003], proved to be correct in the case of brow actions. In brief, the properties of the temporal dynamics of brow actions that help in distinguishing spontaneous from deliberate facial expressions are:

1. maximal displacement of the relevant facial points (relating to the intensity of shown AU),
 2. maximal velocity of this displacement (relating to the speed and the trajectory of AU activation),
- and

AU	onset	apex	offset
1	0.797	0.703	0.843
2	0.833	0.532	0.863
4	0.751	0.694	0.789

Table 9.2: Correct classification rates attained by 9 RVMs using the mid-level feature parameters that the Gentle Boost selected as the most informative for the classification problem in hand, i.e., determining whether a detected temporal segment of an activated AU has been displayed spontaneously or not.

Classification rate	0.908	Total spontaneous:	70
Recall	0.829	Correct spontaneous:	58
Precision	0.921	Total deliberate:	114
F1-measure	0.873	Correct Deliberate:	109

Table 9.3: Final classification results achieved by the probabilistic decision function defined by equation (9.13) for the entire brow actions shown in an input face image sequence. For the purposes of computing the recall and precision, the spontaneous class was considered the target class.

3. the occurrence order of the action within the image sequence (relating to co-occurrences of AUs).

The symmetry of facial actions, reported to be an important parameter for distinguishing spontaneous from posed smiles did not appear to be important in the case of brow action. That at least must be our conclusion given our data and the methods we have used to extract information from it about the brow action symmetry.

Table 9.2 summarises the results of evaluating the performance of 9 GentleRVMs using the leave-one-subject-out cross validation. These results can be read as follows: the higher the recognition rate for a certain temporal segment, the more informative the segment is for distinguishing between spontaneous and deliberate brow actions. Hence, onset and offset phases are very important in distinguishing voluntary from involuntary brow actions while the apex phase does not contribute much (or not at all in the case of AU2) to the process. This finding reconfirms and reinforces the observation that temporal dynamics of facial actions play a crucial role in distinguishing spontaneous from deliberate facial displays given that the apex phase of a facial action can be regarded as static in the sense that the shape and appearance of a face do not change during this phase.

Table 9.3 summarises the final classification results achieved by the probabilistic decision function defined by equation (9.13). These results clearly indicate that the proposed method is effective for distinguishing voluntary from involuntary brow actions.

It is interesting to note that the achieved correct classification rate is much higher (90.8%) for the event coding where the event is the entire brow action displayed in the input video than for the

event coding where the event is one of the temporal segments of one of the displayed brow actions (75.6% on average, from Table 9.2). This reconfirms once again the finding that temporal aspects of facial displays are important in distinguishing voluntary from involuntary facial expressions. More specifically, the observed difference in the achieved correct classification rates can be read as follows. Even brief observations of facial behaviour are sufficient to make rather accurate judgements on deliberateness of the shown facial signals. When observations are longer, cues at a higher semantic level become available, revealing temporal relations between isolated cues extracted from fleeting glimpses of behaviour and leading to greater prediction accuracy.

This reminds (partially) of observations made in the study on thin slices of behaviour, which suggests that human judgements of very short observations of behaviour may be less accurate but that for observations lasting 30 seconds or longer, the judgements are accurate and do not become significantly more accurate as the length of the observations increases [Ambady and Rosenthal, 1992]. We could not confirm or reject the latter assumption for the case of judgements made by a computer system rather than by a human observer, since the examples of facial behaviour used in our study were short, lasting 5 to 35 seconds.

We also measured the effect of incorporating confidence measures p_t into the final decision function defined by equation (9.12). To do so, we evaluated the performance of the proposed method using a redefined final decision function such that (9.12) is redefined as $\gamma = \sum_{t=1}^m c_t$ to leave out the confidence. The attained correct classification rate then dropped to 87.0%, a decrease of almost 4% in correspondence to the correct classification rate realized when using the originally proposed final decision function. This clearly shows the benefit of using a probabilistic decision function rather than a deterministic final decision function.

9.3 Posed vs. spontaneous smiles: multi-cue automatic detection

Psychological research findings suggest that humans rely on the combined visual channels of face and body more than any other channel when they make judgements about human communicative behaviour [Ambady and Rosenthal, 1992]. However, as discussed in chapter 2, most of the existing expression analysers are monomodal and target human facial affect analysis by learning to recognise a small set of prototypical emotional facial expressions such as happiness and anger [Pantic and Rothkrantz, 2003, Pantic and Bartlett, 2007]. Exceptions from this overall state of the art that do treat facial and bodily cues together include works that combine facial and bodily cues [Gunes and Piccardi, 2007] or facial, bodily and vocal cues [Karpouzis et al., 2007] to detect emotions. Works that attempt to detect non-

prototypical facial expressions include few tentative efforts to detect attitudinal and non-basic affective states such as fatigue [Gu and Ji, 2004] or pain [Bartlett et al., 2006] from face video. As discussed in chapter 2, there exist a number of promising systems that are capable of detecting 15 to 27 AUs.

In addition, independently of whether the approach is AU- or affect-oriented, most of the past work on automatic facial expression analysis is aimed at the analysis of posed (i.e., volitionally displayed) facial expression data. The short list of papers described in section 9.1 is practically exhaustive, only a number of conference papers whose contents were covered by the same authors in a follow-up journal paper and two paper by the author of this thesis were omitted. However, to the best of our knowledge, no vision-based system exists yet that is based on FACS, takes multiple behavioural cues into account (e.g., facial, head and shoulder gestures), and automatically discerns between posed and spontaneous expressions.

Also, none of the systems discussed in section 9.1, which can handle spontaneous data, take the temporal dynamics of facial actions into account, not in the computation of their features nor in the design of their classification procedures. The notable exception is the first work on spontaneous expressions [Bartlett et al., 2003], that uses Hidden Markov Models for AU activation detection. Some of the works described in section 9.1 (e.g. [Littlewort et al., 2007]) do acknowledge cognitive scientists' findings which indicated that the temporal behaviour of facial expressions is of paramount importance to human action understanding, yet they chose to ignore that aspect of the problem. Our proposed method, on the other hand, focuses explicitly on the temporal dynamic aspects of facial expressions.

Overall, computer vision and human-computer interaction (HCI) communities have not adequately exploited the expressive information carried by the body modality [Hudlicka, 2003]. Attempts to recognise affective body movements are few and efforts are mostly on the analysis of posed body actions without considering the facial actions (e.g., [Camurri et al., 2005]). Static postures of acted emotions were recorded by De Silva et al. [Silva et al., 2005] using a motion capture system, but the authors did not attempt to recognise the postures automatically.

Although it has been commonly stated that reliable assessment of human affect requires the concurrent use of multiple cues or modalities [Hudlicka, 2003], relatively few works have focused on implementing affect recognition systems using multi-cue or multimodal data [Pantic et al., 2005a]. Gunes and Piccardi [Gunes and Piccardi, 2007] presented an approach to bimodal recognition of posed expressions of emotions by recording face and body gestures simultaneously using two cameras. Kapoor and Picard focused on the problem of detecting the affective states of high-interest, low-interest and refreshing in a child who is solving a puzzle [Kapoor and Picard, 2005]. They combined sensory information from the face video, the posture sensor (a chair sensor) and the game being played in a

probabilistic framework. Karpouzis et al. [Karpouzis et al., 2007] fused data from facial, bodily and vocal cues using a simple recurrent network to detect emotions. Cohn et al. [Cohn et al., 2004a] conducted a multi-cue analysis of spontaneous smiles, comparing data from the face and the head modalities. However, in that study no fusion of the data was carried out. The work represents a behavioural science study rather than an attempt to automate human affect recognition from multiple visual cues. The study showed that there exists a correlation between the two modalities. The major findings reported by these works were: when recognising affective states from multimodal non-verbal data, body gestures or postures provide better information than other cues (i.e., face), and the fusion of multiple cues and modalities significantly outperforms classification using the individual modalities [Gunes and Piccardi, 2007, Kapoor and Picard, 2005].

In this section we propose a method for automatic, multi-cue, vision-based discrimination between posed and spontaneous smiles. We focus on smiles because of their importance in human development and communication. Developmentally, smiles are one of the first emotion expressions to appear, they occur with relatively high frequency throughout the lifespan and they express a multitude of meanings, including joy, appeasement and greetings, and they often serve to mask anger, disgust, and other negative emotions [Cohn and Schmidt, 2004].

In this study, we explore the following three issues: Firstly, we want to know what the relative importance of the face, the head and the shoulders are for the problem of posed vs. spontaneous smile recognition. It is widely accepted that facial expressions reveal whether a display of affect is posed or genuine [Cohn and Schmidt, 2004, Ekman, 2003, Hess and Kleck, 1990]. However, there is no such consensus when it comes to the relevance of bodily motion.

Darwin argued that because our bodily actions are easier to control consciously than our facial expressions, the information contained in the signal of body movements should be less significant than that contained in the facial expressions, at least when it comes to discerning spontaneous from posed behaviour [Ekman, 2003]. Ekman however, argued that people do not bother to censor their body movements [Ekman, 2003] and therefore, the body would be the more *'leaky'* source. Ekman also reported that a third of his subjects showed a fragment of a shrug when lying [Ekman, 2003]. Furthermore, research in nonverbal behaviour and communication theory stated that deceptive behaviour differs from truthful behaviour in that it lacks head movement [Buller et al., 1994] and illustrating gestures which accompany speech [DePaulo, 2003, Vrij et al., 2000]. Distinguishing between posed and spontaneous smiles can be seen as similar to the deception detection problem. Taking into account these observations, we expect to find valuable information concerning the nature of a nonverbal expression (i.e., posed or spontaneous) in head and shoulder movements as well as in facial actions.

Secondly, we want to investigate the importance of the temporal dynamics of human nonverbal behaviour within the problem of posed vs. spontaneous smile recognition. In section 9.2.1 we already listed the research done by cognitive scientists when it comes to the importance of temporal dynamics of facial actions. For the body modality, DePaulo et al. reported that deceivers' body actions appeared overcontrolled and abrupt [DePaulo, 2003]. Based on these findings, we focus on both the morphological aspect (e.g., AU12, shoulder shrug, posture of the head, etc.) and on the temporal aspect (e.g., speed, trajectory etc.) of the expressive face and body display and we expect that they will play a significant role in the recognition of posed vs. spontaneous smiles.

Thirdly, we want to look into the effect that different multi-cue data fusion strategies have on the classification accuracy of posed and spontaneous smiles and compare these with monomodal classification results. The fusion strategies differ from each other mainly in two aspects: the abstraction level of the utilised features and the way in which the classification results are combined. Regarding the level of abstraction, it is widely argued that feature-level fusion outperforms decision-level fusion in the field of non-verbal human behaviour analysis [Pantic and Rothkrantz, 2003].

Designing optimal strategies for realising multi-cue and/or multi-cue data fusion is still an open research issue. Various approaches have been proposed [Fumera and Roli, 2005, Kuncheva, 2002] including the sum rule, product rule, weight-based fusion, maximum/minimum/median rule, majority vote, etc. The first three techniques, namely, the sum, product and weight criteria were found to be most effective and are analysed and compared here for this problem.

9.3.1 Tracking

We employ a different tracker for each modality: a Cylindrical Head Tracker to track the head motion [Xiao et al., 2003], Particle Filtering with Factorised Likelihoods to track 20 fiducial points in the face (see section 5.1, [Patras and Pantic, 2004, Patras and Pantic, 2005]), and Auxiliary Particle Filtering to track the shoulders motion [Pitt and Shephard, 1999]. From the 20 tracked facial points we will only use 12 for this study. Fig. 9.2 shows the facial and shoulder points that we will use.

To capture the head motion we employ the Cylindrical Head Tracker developed by Xiao et al. [Xiao et al., 2003]. The head tracker estimates the six degrees of freedom of head motion: horizontal and vertical position in the scene, distance to the camera (i.e. scale), pitch, yaw and roll. This is denoted as the set of parameters $T_h = \{T_{h1} \dots T_{h6}\}$ with dimensions $n \times 6$. Here n is the number of frames of the input image. A cylindrical head model is manually fitted to the face region in the first frame (see Fig. 9.1), and the face image is cropped and projected onto the cylinder as the

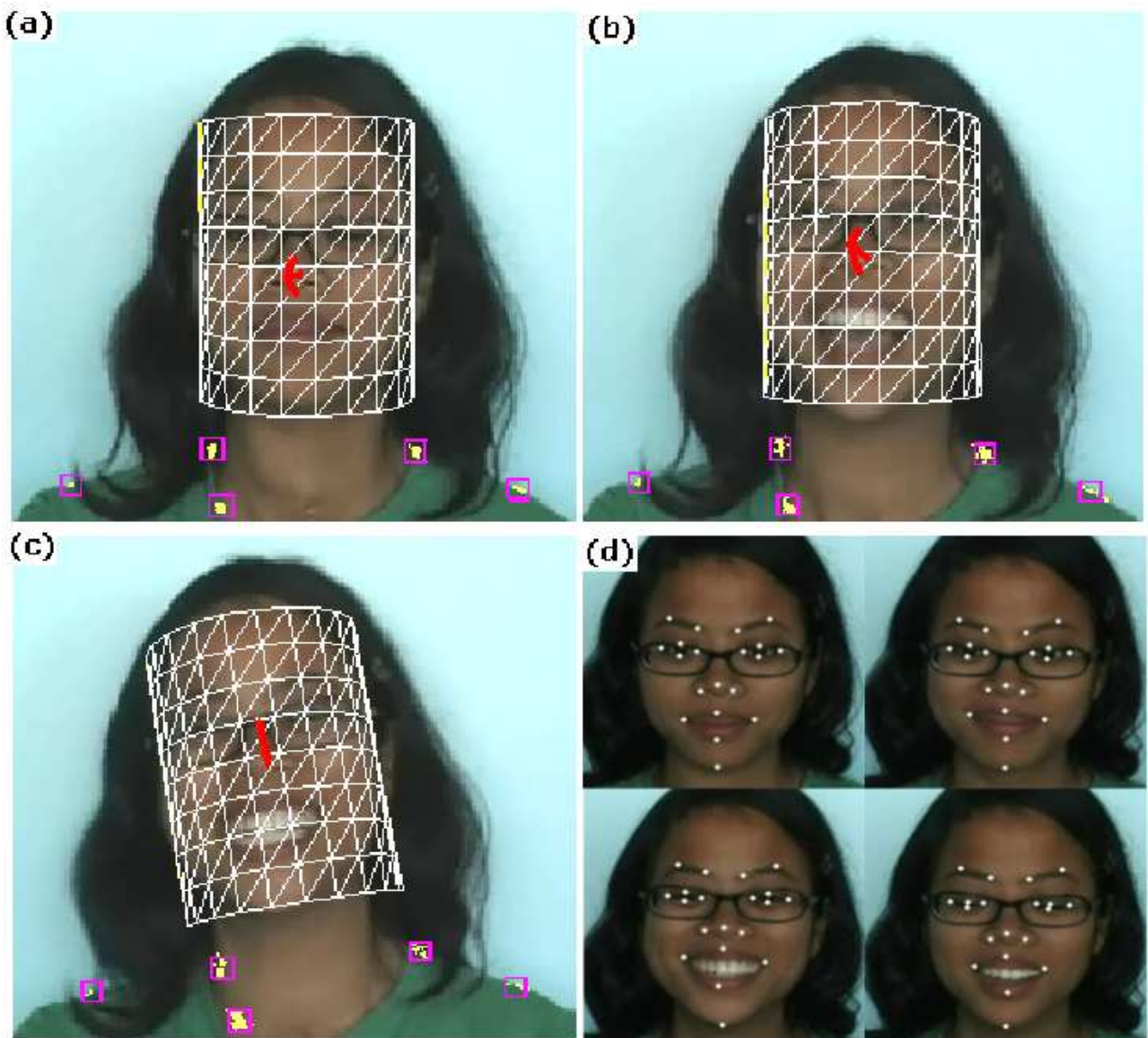


Figure 9.1: Illustration of the tracking procedure and the points used to obtain the tracking data: (a-c) for the head and shoulder modalities, and (d) for the face modality.

template of head appearance. For any given subsequent frame, the face template is projected onto the image plane assuming that the pose has remained unchanged from the previous frame. Then, the difference between the projected image and the current frame is computed, providing the correction of the estimated head pose. Because the templates are updated during head motion recovery, the errors of motion recovery accumulate over time. To tackle this problem, images of certain reference poses are prepared and when the estimated head pose is close to that in the reference, the head image is re-registered with the relevant reference image. The re-registration also enables the system to recover the head pose when the head reappears after occlusion.

To capture all facial motion that is characteristic for smiles, we first track the 20 facial points as

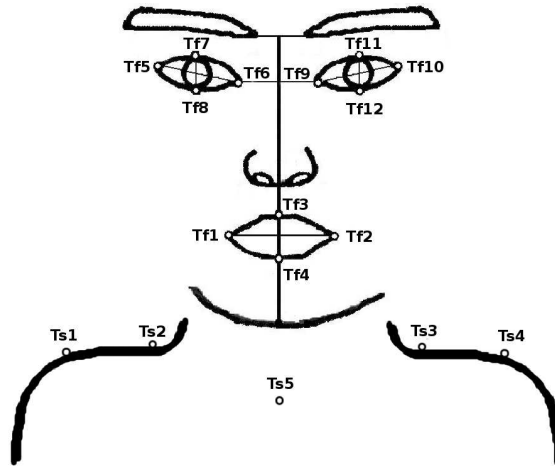


Figure 9.2: Tracked points $T_{f1} \dots T_{f12}$ of the face and tracked points $T_{s1} \dots T_{s5}$ of the shoulders.

described in section 5.1. From this set of points we select 12 points that we will use for this study. From the mouth we use the mouth corners, the upper, and the lower lip. From the eyes we use the inner and outer eye corners as well as the upper and lower eyelids. For this study, the facial points were manually localised in the first frame of every video. We used the head tracker data to register the images so that the face and its features (the facial points) have the same frontal position in every frame. After selection of the points the facial point tracking scheme results for every image sequence in a set of tracked points $T_f = \{T_{f1} \dots T_{f12}\}$ with dimensions $n \times 12 \times 2$.

The motion of the shoulders is captured by tracking 2 points on each shoulder and one stable point on the torso, usually just below the neck (see Fig. 9.2. Often we chose a salient point on the clothes that subjects wore). The stable point is used to remove any rigid motion of the torso (see section 9.3.3). We use standard Auxiliary Particle Filtering (APF) [Pitt and Shephard, 1999] instead of PFFL because it is less complex and faster than PFFL, it does not require learning the model of prior probabilities of the relative positions of the shoulder points, while resulting in sufficiently high accuracy. The shoulder tracker results in a set of tracked points $T_s = \{T_{s1} \dots T_{s5}\}$ with dimensions of $n \times 5 \times 2$.

9.3.2 Temporal segmentation

In chapter 7 we described how a facial muscle action can be divided into four phases. Just like facial actions, head and body actions can be in only one of the four possible phases at any time; the onset phase, the apex phase, the offset phase, and (iv) the neutral phase. According to [Cohn and Schmidt, 2004, Ekman, 2003] and reaffirmed in section 9.2, timing, duration and speed of facial actions are highly important cues for distinguishing posed from spontaneous facial expressions. Especially the speed at

which a facial expression develops or diminishes during the onset and offset phases respectively has proved to be highly discriminative. In section 9.3.3 we will define these cues in terms of the speed of tracked points or, in the case of the head, the angular and translational velocity of the entire head. In order to be able to compute these cues, or attributes, separately for the onset, apex and offset temporal segments of a face/body action, we first need to know when every temporal segment of every modality begins and ends.

For the recognition of the temporal segments of the face, we use the method proposed in chapter 7. We detect temporal segments for AU6, AU12 and AU13. This segmentation results in a number of $m_f = m_{AU6} + m_{AU12} + m_{AU13}$ temporal segments for the face modality.

The temporal segments of the head and the shoulder modalities are obtained using a rule-based expert system as we do not have the manually labelled temporal segment data for these modalities based on which an automatic detector could be trained. The temporal segments are found as follows. We only consider one action to be possible for the head and the shoulders, that is we only check if the head or the shoulders are in their neutral position or not. We define:

$$q_{h1}(t) = |\delta(h_1, t)| \quad (9.15)$$

$$r_{h1}(t) = \left| \frac{dh_1(t)}{dt} \right| \quad (9.16)$$

$$q_{h2}(t) = |\delta(h_2, t)| \quad (9.17)$$

$$r_{h2}(t) = \left| \frac{dh_2(t)}{dt} \right| \quad (9.18)$$

where $h_1 = \{T_{h4}, T_{h5}, T_{h6}\}$ is the time series of pitch, roll and yaw and $h_2 = \{T_{h1}, T_{h2}, T_{h3}\}$ is the vector of the head positions. Both are subsets of T_h . Now, for each time t we say that the head is in its neutral phase if both the angle difference and the head translation are close to zero, that is, IFF (if and only if) $q_{h1} < \theta_1$ AND $q_{h2} < \theta_2$. If the head is not in its neutral phase, we continue to check whether the head is in its apex phase, that is, whether the angular velocity or the translational velocity are sufficiently close to zero. This is the case IFF $r_{h1} < \theta_3$ AND $r_{h2} < \theta_4$. If the head is neither in its neutral nor in its apex phase, we check the sign of the first derivative to time of the angular motion $d(\delta(r_{h1}, t))/dt$ and the current position of the head. If the head is moving away from its neutral position, we assign the onset phase, otherwise we assign the offset phase. This results in m_h temporal segments for the head modality.

Similarly, for the shoulders, we find

$$d_{s1}(t) = |\delta(s_1, t)| \quad (9.19)$$

$$m_{s1}(t) = \left| \frac{ds_1(t)}{dt} \right| \quad (9.20)$$

$$d_{s2}(t) = |\delta(s_2, t)| \quad (9.21)$$

$$m_{s2}(t) = \left| \frac{ds_2(t)}{dt} \right| \quad (9.22)$$

where $s_1(t)$ denotes the angle made by the horizontal axis and the line connecting shoulder points T_{s1} and T_{s2} at time t and $s_2(t)$ is the angle made by the horizontal axis and the line connecting shoulder points T_{s3} and T_{s4} at time t . The temporal phases of the shoulder actions are now found in the same way as described above for the head, resulting in four more thresholds $\theta_5 \dots \theta_8$. This way we find m_s temporal segments for the shoulder modality, and in total $m = m_f + m_h + m_s$ temporal segments for all modalities together. The values of the thresholds $\theta_1 \dots \theta_8$ were found using cross validation during training. This way we avoid overfitting of the threshold values to our data.

9.3.3 Fusion strategies

One of the goals of this study is to investigate the effects on the recognition accuracy of using different levels of abstraction at which features are defined, different abstraction levels of the classification schemes (i.e. only one classifier working directly on features, or multiple classifiers working in layers after each other), and different fusion rules. In order to achieve this, we implement three different fusion strategies to tackle the problem of multi-cue posed vs. spontaneous smile recognition: early, mid-level and late fusion. We distinguish two levels of feature abstraction here. Early fusion uses low-abstraction features while mid- and late level fusion use high-abstraction features. In the next subsections, we will describe each fusion strategy in detail, including the definition of the attributes used for each.

Early fusion

In early (also called feature-level) fusion, the elementary attributes of each visual cue are combined into a single low-abstraction feature vector, which serves as the input to one classifier. In our case, the elementary attributes are simple operations on the tracking data, such as the distances between two tracked points or the current angular velocity of the pitch of the head. All the features are computed

and passed to the classifier on a per-frame basis, resulting in a time-series of classification predictions \mathbf{y} with length n for every input video.

From the head modality we simply used the output from the tracker (see section 9.3.1):

$$f_h = T_h \quad (9.23)$$

From the face modality, we concatenated the x - and y -values of all tracked points, the distances between all pairs of points and the angles between the line connecting two points and the horizontal axis. Thus we create for the face the following feature vector $f_f(t)$:

$$f_f(t) = \{f_1(T_{f1}, t), \dots, f_1(T_{f12}, t), \quad (9.24)$$

$$f_2(T_{f1}, t), \dots, f_2(T_{f12}, t), \quad (9.25)$$

$$f_3(T_{f1}, T_{f2}, t), \dots, f_3(T_{f11}, T_{f12}, t), \quad (9.26)$$

$$f_4(T_{f1}, T_{f2}, t), \dots, f_4(T_{f11}, T_{f12}, t)\} \quad (9.27)$$

where $T_{f_i}(t)$ is the value of tracked facial point i at time t . See section 5.3 for the definitions of the geometry-based features. For the shoulder modality, we defined the feature vector as follows:

$$f_s(t) = \{f_4(T_{s1}, T_{s2}, t), f_4(T_{s4}, T_{s3}, t), [\delta(T_{s1,y}, t) + \delta(T_{s2,y}, t)], [\delta(T_{s3,y}, t) + \delta(T_{s4,y}, t)]\} \quad (9.28)$$

See Fig. 9.2 for the numbering of the shoulder points. To obtain our final feature vector used in early fusion, we first concatenated the above attributes as $f = \{f_h, f_f, f_s\}$. Then, we defined

$$F_e(t) = \{f(t), df(t)/dt, \delta(f, t)\} \quad (9.29)$$

In this definition of our features, we denominated the first term of the right hand side of eq.(9.29) as the static features and the second and third terms as the dynamic features.

The feature vector $F_e(t)$ serves as input to a GentleSVM-Sigmoid classifier. This classifier performs feature selection using GentleBoost and classification using SVMs, as discussed in chapter 6. Unfortunately the output of an SVM is not a good measure for the posterior probability of its prediction. Therefore we pass the output of the SVM to a sigmoid function that has been shown to provide a

reasonable measure for the posterior probability [Platt, 2000]. We will refer to this feature selection-classifier combination as a GentleSVM-Sigmoid classifier. The vector $F_e(t)$ provides the t^{th} element of the time series of predictions, $\mathbf{y}(t)$. Under a Maximum-a-Posteriori (MAP) approach, $\mathbf{y}(t)$ must be assigned to one of the two possible classes (posed or spontaneous), according to maximum posterior probability.

Mid-level fusion

Mid-level fusion attains a higher level of data abstraction within every visual cue, yet we still fuse all attributes into one vector, and use only one classifier. Often in mid-level fusion the elementary attributes of early fusion are used to define a set of abstract symbols (such as AUs), or they are used to compute more heuristic features. In our approach, we transform the elementary attributes derived previously into both symbols and higher level features. The symbols we derive at this stage are the temporal segments of the FACS Action Units AU6, AU12 and AU13 [Ekman et al., 2002] and the temporal segments of the head and shoulder action. We refer to the combination of symbols and the higher level features that describe timing aspects of the symbols as the high-abstraction features.

For each temporal segment, we define attributes based on the works of Cohn and Schmidt, Ekman, and on the method proposed in section 9.2, [Cohn and Schmidt, 2004, Ekman, 2003]. These include the morphology, speed, symmetry, duration, apex overlap (i.e., number of frames that two actions are in apex simultaneously), trajectory of a face/body action and the order in which the temporal segment of the different face/body actions occur. Similar to the case of early fusion, we concatenate the attributes of all visual cues into one vector, which serves as the input to a GentleSVM-Sigmoid classifier. In contrast with early fusion, which computed one feature vector per frame, we now compute one vector per video, with every temporal segment of every symbol providing a fixed number of attributes.

Because the GentleSVM-Sigmoid classifier requires the input vectors to be of the same length, we were forced to change our definition of the temporal segments. Face/body actions (especially when displayed spontaneously) frequently have multiple apices before returning to its neutral position. This is especially the case for spontaneous smiles. If left unchanged, this would lead to feature vectors of variable length. We decided therefore to force any sequence of temporal segments into the temporal pattern onset-apex-offset. To do so, we chose the apex segment with the longest duration as the apex phase of our new forced temporal pattern. The frame at which the new — forced — onset starts is chosen as the first non-neutral frame before the apex phase. Conversely, the end of the new forced offset is chosen as the last non-neutral frame after the apex phase. The neutral segments are discarded as, by definition, they do not contain any information. This results in k forced temporal segments for

all visual cues together.

By enforcing this particular temporal pattern, we are aware of the fact that information about the original temporal pattern of the symbols, such as multiple apices, is lost. To be able to still use this information in the classification process, we compute a number of concurrency attributes M for every video. The concurrency feature vector M is computed before we enforce the new temporal pattern and contains the duration of all symbols, the order in which the symbols are activated and the duration of overlap of every combination of apex phase symbols.

We thus define for every temporal segment of the forced pattern the following set of features. For each segment of a face action (either AU6, AU12, or AU13) or shoulder action (motion) we define the mean/max displacement and the mean/max velocity of the tracked properties during that segment as:

$$g_{symbol} = \left\{ \frac{\sum_{t=t_1}^{t_m} \delta(T, t)}{t_m + 1 - t_1}, \max_{t=t_1 \dots t_m} \delta(T, t), \frac{\sum_{t=t_1}^{t_m} d(T(t))/dt}{t_{m2} + 1 - t_1}, \max_{t=t_1 \dots t_m} d(T(t))/dt \right\} \quad (9.30)$$

where t_1 and t_m are the first and the last frames of a temporal segment, respectively. T is the tracking data of the eye points when considering AU6 symbols, the mouth points when considering AU12 or AU13 symbols, the head tracking when considering head action symbols, and the shoulder tracking data when considering shoulder action symbols. Additionally, for the face and shoulder cues we compute the asymmetry value a . For the head cue, we define a to be always zero, as there is no such thing as a symmetrical head action. The final feature vector for mid-level fusion is thus found as the union of M , g and a for all symbols S , where S consists of all possible combinations of {AU6,AU12,AU13, head action, shoulder action} and the forced temporal segments {onset, apex, offset}:

$$F_m = \left\{ \bigcup_{i \in S} (g_i), \bigcup_{i \in S} (a_i), M \right\} \quad (9.31)$$

Because the values of all features depend on the time parameters t_1 and t_m we consider all mid-level parameters to be temporal dynamic.

Given the feature vector F_m , where F_m describes an entire smile, we again use a GentleSVM-Sigmoid classifier to predict the class of the video under the previously described MAP approach.

Late fusion

Late fusion is similar to mid-level fusion in the sense that we again attain a high level of data abstraction within every visual cue. However, in the case of late fusion we also attain a higher level of abstraction for the classification procedure, computing a separate posterior probability for each symbol (i.e. for each temporal phase of each AU, shoulder action and head action). This removes the need to enforce the strict onset-apex-offset temporal pattern used in mid-level fusion. Indeed, we will use all m temporal segments, collected from all visual cues. In this way we obtain a variable number of predictions y for every video. This enables the system to discard a cue when needed (e.g., when the shoulders move out of view), as the fusion rules we employ are invariant to the number of inputs. We compute one extra posterior probability that encodes temporal dynamic relations between the various signals. This posterior is computed based on a concurrency feature vector M that contains the duration of all temporal segments for each visual cue, the order in which the segments are activated and the number of frames that the apex phases overlap of for every combination of symbols.

Every temporal segment from every symbol generates one feature vector. This vector is defined as:

$$F_l(i) = \{g(i), a(i)\} \quad (9.32)$$

where i is a temporal segment. Again, all features are considered dynamic. Each feature vector $F_l(i)$ is used as input to the appropriate GentleSVM-Sigmoid classifier. That is, we train a different classifier for every temporal segment type for every symbol. Thus we train one GentleSVM-Sigmoid for the onset phase of AU12, one for the apex phase of shoulder actions, etc. A separate classifier is trained for the concurrency attributes. The vector $F_l(i)$ then provides the i^{th} element of predictions y . After achieving this, the general approach of late fusion of the individual classifier outputs can be described as follows.

The time series y represents the whole image sequence and $F_l = (F_{fl}, F_{hl}, F_{sl})$ represent the overall feature vectors consisting of the face F_{fl} , head F_{hl} , and shoulder F_{sl} feature vectors. Under a Maximum-a-Posteriori (MAP) approach, y must be assigned to one of the two classes (w_1, w_2), having maximum posterior probability $p(w_k|y)$. Once the posterior probabilities per visual cue, per temporal segment are obtained by again passing the output of the SVM to a sigmoid function as proposed by Platt [Platt, 2000], late fusion is applied. The three separate classifiers provide the posterior probabilities $p(w_k|F_{hl})$, $p(w_k|F_{fl})$ and $p(w_k|F_{sl})$ for the head, face and shoulder cues, respectively, to be combined into a single posterior probability $p(w_k|y)$ with one of the fusion methods described in Table 9.4. Note that the weights are derived from the classification results on the training data during

Table 9.4: Description of the three late fusion criteria used: sum, product and weight.

sum	$k = \operatorname{argmax}_{k=1}^2 p(w_k F_{fl} + F_{hl} + F_{sl})$
product	$k = \operatorname{argmax}_{k=1}^2 p(w_k F_{fl} \times F_{hl} \times F_{sl})$
weight	$k = \operatorname{argmax}_{k=1}^2 \sigma_f p(w_k F_f) + \sigma_h p(w_k F_{hl}) + \sigma_s p(w_k F_{sl})$

Table 9.5: Selected low-abstraction features to distinguish posed from spontaneous smiles.

Relevance	Modality	Feature definition
1	Face:	$\delta(f_3(T_{f3}, T_{f4}))$
2	Shoulders:	$f_2(T_{f2})$
3	Face:	$\delta(f_4(T_{f1}, T_{f4}))$
4	Head:	$f_1(T_{f3})$
5	Face:	$\delta(f_2(T_{f2}))$
6	Face:	$f_4(T_{f2}, T_{f4})$
7	Shoulders:	$f_2(T_{f4})$
8	Face:	$f_4(T_{f3}, T_{f4})$
9	Shoulders:	$f_1(T_{f3})$
10	Head:	$f_2(T_{f1})$
11	Shoulders:	$f_1(T_{f1})$
12	Head:	$f_3(T_{f1}, T_{f4})$
13	Face:	$\delta(f_4(T_{f1}, T_{f3}))$
14	Shoulders:	$f_1(T_{f2})$
15	Face:	$\delta(f_1(T_{f4}))$

cross validation.

9.3.4 Evaluation

We evaluated the three fusion approaches on 100 videos of posed smiles and 102 videos of spontaneous smiles using 10-fold cross-validation. The videos of posed smiles were taken from the publicly available part of the MMI database. The videos of spontaneous smiles were taken from three different databases: 8 videos from the Triad database [Kirchner et al., 2006], 3 recordings from the MMI database, and 91 videos from the MMI-database part 2. For this study, we normalised all measurements using the inter-ocular distance D_I . This way we make sure we do not accidentally learn to recognise an entire database using the average distance of the face to the camera.

All videos were recorded from a near-frontal view, under controlled lighting conditions. The recording of the MMI-database is described in section 3.4. The Triad database was obtained to study the effect of alcohol on the social behaviour of men. Three men sitting around a table were recorded while talking freely to each other and consuming alcohol. All videos were edited to ensure that they contained

exactly one smile. Multiple apices were allowed; the video was only cut when the face had returned to its neutral phase.

The recognition rate of AU detection is important because it influences the calculation of the high-level features used in mid-level fusion and late fusion. If an AU is not correctly detected, temporal segmentation will be flawed and hence the values of the high-level features will be flawed as well. We measured AU recognition rates separately for the spontaneous and the posed expressions. On the spontaneous data, AU6 was recognised correctly with 77% of the times, AU12 with 54% and AU13 with 85%. On the posed data AU6 was recognised correctly 76% of the times, AU12 40% of the times and AU13 76% of the times. The reason why AU12 has a rather low classification rate is that AU12 and AU13 are very similar. Both involve movement of the outer mouth corners. The difference lies in the horizontal movement: with AU12 the mouth corners move further out while with AU13 the mouth corners are pulled up sharply.

Table 9.6 shows the classification results for all fusion strategies. All results were obtained using 10-fold cross-validation. For the purpose of computing the precision and recall, we considered spontaneous smiles to be the positive class. Overall we can say that the proposed system works as desired, being able to discern between posed and spontaneous smiles with fairly high accuracy. There is no significant difference between early and mid-level fusion at a 5% significance level. Late fusion does score significantly higher than early fusion and mid-level fusion. The results also show that even though some AUs were not recognised perfectly, the influence of the low AU detection rates on distinguishing posed from spontaneous smiles is hardly measurable. As the AU detection only has influence on the mid-level and late fusion strategies, we could expect an even higher score for these approaches if the AU recognition would improve.

The high results for late fusion could be explained by two factors. First there is the high classification abstraction. Specialised classifiers are learnt to distinguish posed from spontaneous smiles for each segment of a smile, i.e., during the onset of head motion, the apex of a smile, etc. Because all specialised classifiers return a posterior probability, the fusion rule can then be used to generalise from the results per segment of an action to the entire action (i.e. a smiling face with its accompanying bodily action). Of course this setup is unable to learn which bodily actions typically co-occur with certain facial actions.

The second explanation for the high score for late fusion is the high data abstraction. The low-abstraction features only describe simple attributes: positions, distances and angles of points. Moreover, they only describe those attributes at one point in time — the frame for which they are defined. The high-abstraction features capture more general physical phenomena such as the duration of a

Fusion strategy	Cl. rate	Recall	Precision	F1-measure
Early	0.886	0.889	0.880	0.885
Mid-level	0.881	0.883	0.886	0.885
Late (sum)	0.931	0.956	0.920	0.937
Late (product)	0.940	0.964	0.933	0.948
Late (weight)	0.931	0.943	0.927	0.935

Table 9.6: Classification, recall, precision rates and F1-measure for the different fusion strategies employed. Performance measures are computed per video.

temporal segment, the average speed during onset and the order in which actions occur. One could argue that if the higher data abstraction would benefit the classification procedure, we should also have obtained better results for mid-level fusion than presented in this study. However, for mid-level fusion to work, the original temporal phase pattern, with possibly multiple apices, had to be cast into a strict neutral-onset-apex-offset-neutral phase transition sequence (see section 9.3.3). Because of these classification procedure constraints, the temporal segmentation of the face and body actions was severely altered. We think that this might have cancelled out all the benefits of the high-abstraction features.

To the best of our knowledge, the system presented here is the first to propose discerning posed from spontaneous smiles by fusing video data from face, head and body actions. Therefore we cannot compare our results with other works. Yet, to give an indication of how difficult the problem is, Hess et al. [Hess et al., 1989] achieved an 82% classification rate on a set of 80 smiles using electromyography (EMG).

To investigate what the relative importance of each visual cue was, we investigated the properties of the early fusion strategy. We performed seven tests (again using 10-fold cross-validation), each time using a different combination of visual cues. So in the first three tests we use features from early fusion that belong to only one of the three cues and in the next three tests we use a combination of two visual cues and the last test used all three cues. The results of this test are shown in Table 9.7. For enhanced resolution, the results listed are on a per-frame basis, instead of per-video (so the classification rate indicates the fraction of correctly classified frames). In addition, Table 9.8 provides a matrix showing which of the results were statistically different on a 5% significance level.

When we only take single visual cues into account (combinations I, II and III), the head cue performs best according to our results. However, the recognition rates between the separate visual cues are not significantly different at a 5% significance level ($P = 0.05$). Early fusion of all visual cues (combination VII) *is* significantly better than any of the single-cue combinations. Table 9.7 also shows that a combination of two visual cues already outperforms the tests with only one cue.

Visual Cues	Cl. rate	Recall	Precision	F1-measure
I Face	0.812	0.841	0.868	0.854
II Head	0.822	0.823	0.916	0.867
III Shoulders	0.794	0.793	0.915	0.850
IV Face-Head	0.867	0.897	0.893	0.895
V Face-Shoulders	0.871	0.896	0.899	0.898
VI Head-Shoulders	0.845	0.861	0.899	0.880
VII All	0.895	0.919	0.916	0.918

Table 9.7: Comparison of performance measures for the different visual cues separately and fused. Performance measures are computed per frame.

	I	II	III	IV	V	VI	VII
I	0	0	0	0	1	0	1
II	0	0	0	0	1	0	1
III	0	0	0	1	1	1	1
IV	0	0	1	0	0	0	0
V	1	1	1	0	0	0	0
VI	0	0	1	0	0	0	1
VII	1	1	1	0	0	1	0

Table 9.8: Matrix of statistical significant different classification rates. Roman indices relate to the visual cue combinations listed in table 9.7. A 1 indicates statistically significantly different results.

To further investigate the relevance of the different visual cues, we performed an analysis of the features selected by GentleBoost. For early fusion, 62% of the selected features originated from the face cue, 16% from the head cue, and 22% originated from the shoulder cue. For mid-level fusion, 40% of the features came from the face cue, 40% from the head cue, 13.3% from the shoulder cue and 6.7% of the selected features originated from the concurrency features (which span all visual cues). For early fusion GentleBoost selected 45 low-abstraction features, of which the first 15 are listed in Table 9.5. All 15 selected features for mid-level fusion are listed in Table 9.10.

In late fusion, feature selection takes place in the separate classifiers specialised for each temporal phase of each visual cue so a comparison of the selected features is not feasible. However, we might learn something from the classification performance of the specialised classifiers for each symbol. Table 9.9 shows the classification results attained when we use only one specialised classifier to classify an entire video into a posed or a spontaneous smile. From this table, we can read two things: first, the head cue seems to be most reliable in late fusion. Second, the offset phase seems to carry the least information.

Based on the results for feature selection and the results of combinations of visual cues, given our data, we might conclude that the head is the most important visual cue for distinguishing between posed and spontaneous smiles. This is in agreement with both other HCI works [Gunes and Piccardi, 2007, Kapoor and Picard, 2005] and cognitive scientists' works [Buller et al., 1994, Ekman and Friesen, 1969].

	Onset	Apex	Offset
Face	0.719	0.612	0.451
Head	0.781	0.826	0.742
Shoulders	0.752	0.766	0.638
Concurrency	0.781		

Table 9.9: Classification rates for the specialised classifiers used in late fusion. There is no temporal phase associated with the concurrency classifier.

But more importantly, the results show that the visual cues complement each other. The result of all cues combined (the fused result, same as the early-fusion result of table 9.6) is significantly better at $P = 0.05$ than any of the other visual cues separately. This confirms our hypothesis that a multi-cue approach benefits posed vs. spontaneous smile detection.

To answer our question regarding the relevance of temporal dynamics for automatic multi-cue posed vs. spontaneous smile recognition, we again provide an analysis of the feature selection process. Table 9.5 shows all the selected features for low-abstraction features and Table 9.10 those for high-abstraction features, including the visual cue that the feature originated from. In the case of high-abstraction features, the originating temporal segment is listed as well. For early fusion, 24.4% of the selected features were static features, while 75.6% of the features were dynamic features.

While the fraction of static features was greater than we expected, we can still clearly see that the temporal dynamics are the most important features for automatic multi-cue posed vs. spontaneous smile recognition. This is also reflected in the high classification results of late fusion, which uses only temporal dynamics. A closer look at Table 9.10 reveals that the speed with which the head rotates is of paramount importance, as is the head’s translation and the sequence in which the temporal segments are ordered. Fig. 9.3 shows a scatter plot of the mean velocity of the right mouth corner in x-direction during onset and offset. As we can see, spontaneous smiles have both a slower onset and a slower offset, consistent with the cognitive sciences’ findings [Cohn and Schmidt, 2004].

Ekman predicted that the asymmetry of facial actions is an indicator for distinguishing posed from spontaneous expressions [Ekman, 2003]. We did not find any evidence for this however, only one of the selected high-level features was an asymmetry feature. Although this observation is in disagreement with Ekman’s findings, the same lack of correlation between asymmetry and the nature of the expression was previously reported by Schmidt et al. [Schmidt et al., 2006].

The results show that the proposed multi-cue approach to automatic distinction between posed and spontaneous smiles is extremely accurate. From the results presented, it is clear that fusing video data from the face, head and shoulders increases the accuracy compared to a monomodal/mono-cue

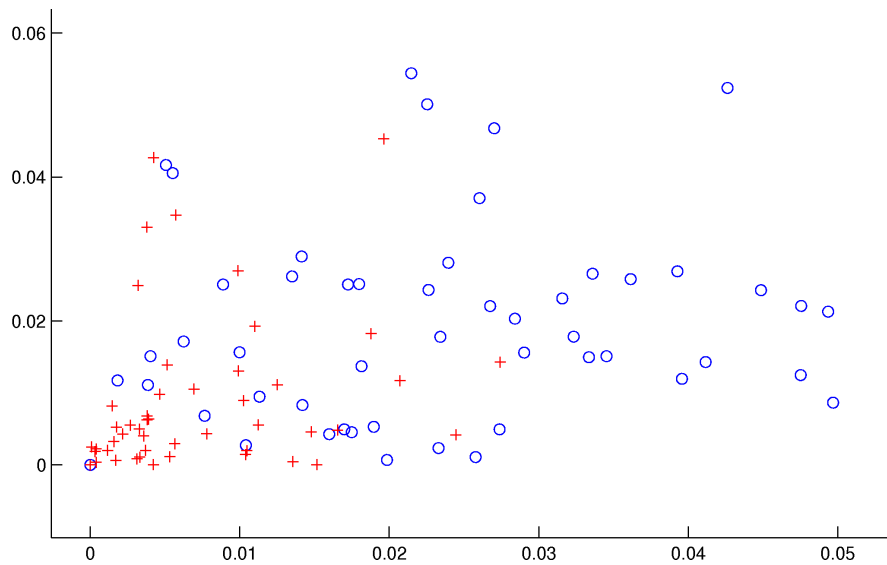


Figure 9.3: Mean velocity of the right mouth corner in x-direction during onset (x-axis) vs. mean velocity of the right mouth corner in x-direction during offset (y-axis). Crosses denote spontaneous smiles.

Rel.	Cue./seg.	Feature
1	onset head:	mean angular velocity
2	concurrency:	order apex shoulders
3	apex head:	max translational displ.
4	apex head:	mean angular velocity
5	apex shoulders:	max angular velocity of left shoulder
6	apex AU6:	Asymmetry in x-direction
7	apex head:	max angular displ.
8	offset AU12:	mean velocity of point 2 in x direction
9	offset AU13:	mean displ. of point 4 in x direction
10	onset shoulders:	max displ. of right shoulder y direction
11	offset head:	mean angular velocity
12	onset AU13:	mean velocity of point 3 in x direction
13	apex AU13:	mean displ. of point 4 in x direction
14	onset head:	max angular velocity
15	apex AU12:	max displ. of point 2 in x direction

Table 9.10: Selected high-abstraction features to distinguish posed from spontaneous smiles.

system that employs only one of the listed visual cues. This is in agreement with the body of work in cognitive sciences indicating that humans leak their intentions not only through facial expressions, but also through their body language. It is hard to say which visual cue is the most important. For the data analysed, results seem to indicate that the head motion is the most reliable source, followed closely by the face. However, more experiments are needed to confirm this. When evaluating the importance of the expression dynamics for classification purposes, the dynamic attributes are clearly more important than static ones. This can be seen from the large number of dynamic features selected during early fusion, as well as from the high results for late fusion (which uses only dynamic features). Regarding the different fusion strategies, late fusion clearly performs best. The first reason for this is that with late fusion we are able to decompose the problem in smaller subproblems, for which we can train specialised classifiers. Another major benefit of late fusion is the use of high-abstraction features, which encode important temporal dynamic attributes of human nonverbal behaviour.

Chapter 10

Conclusion

In this thesis we have investigated the morphologic and temporal dynamic aspects of automatic facial expression recognition from a geometric feature point of view. We applied state-of-the-art face detection, facial point detection and facial point tracking techniques to attain a set of tracked points for a video displaying a facial expression. Based on these tracked points we proposed a geometric feature representation of facial actions, that was shown to encode both information about the morphology of facial actions as well as of their temporal dynamics. A combination of GentleBoost feature selection and SVM classification was used for AU activation detection. For the recognition of the temporal phases of AUs, a hybrid SVM-HMM was proposed, preceded by GentleBoost feature selection. We have shown that these approaches can indeed be very successful for AU analysis, particularly for the analysis of AU temporal dynamics.

This thesis is the first work to address the automatic recognition of the AU temporal phases onset, apex and offset. Not only have we proposed a method that results in high classification accuracy and timing precision for the recognition of these temporal phases, we have also shown that a successful analysis of AU temporal dynamics can help in attaining higher AU event detection results. Moreover, we have shown that we can use the information about the timing aspects of the automatically recognised temporal phases to distinguish between some aspects of human facial behaviour.

10.1 Summary of Thesis Achievements

10.1.1 Feature definitions

This thesis provides more insight in the applicability of geometric features for facial expression analysis than was previously available. We have investigated a number of ways to extract as much information from the facial point tracking data as possible. Three sets of geometric features were defined in chapter 5, each with a different temporal scope. The first set, F_S , uses only the information available at one moment in time, i.e. only the position of points and distances and angles between points in the current frame. Of course it was to be expected that this feature set would not do well to describe temporal patterns in a facial action. Yet somewhat surprisingly, we have shown that this information is not only insufficient for the analysis of the face's temporal dynamics, but also the worst set of features to describe the morphology of a facial action, which could be considered to be a static facial expression recognition problem.

The best set of features was F_D , which computes its features comparing facial point tracking information between two moments in time. The features are defined using the current frame values and either the previous frame (to compute velocities of points) or the first frame in a sequence (to compute deviations with respect to a neutral expression). Strictly speaking, we don't necessarily have to compare with the first frame of a sequence. We could compare with any frame about which we know that the face is neutral in that image. This set achieved the highest classification accuracy, both on its own and in combination with other feature sets.

Somewhat puzzling are the results for the third feature set, F_W . This set was designed to capture more complex local temporal dynamics by describing the evolution of the other features by a polynomial function fitted to the feature values in a fixed time-window. As expected, the added value of this feature set was greatest for detecting the onset and offset of AUs (see table 7.4). However, the set F_D on its own was better at this than F_W alone. Combining F_D and F_W improved performance somewhat, but the full potential was only achieved when F_W was combined with both F_D and F_S . Because of our strict evaluation procedures we are confident that this is not due to overfitting, as might be suggested since the union of all three sets has a greater dimensionality. Yet we cannot fully explain why the descriptive power of the local time parameterised features F_W only come to their right in combination with both the static features and the two-frame based features.

Another important lesson learned from the analysis of the different feature sets is that when we choose to use geometric features, analysis of video will always outperform the analysis of static images. By

definition the only feature set applicable to static images is F_S and we have shown that this results in the worst performance. This holds both for the analysis of the morphological as well as for the temporal dynamic aspects of facial expressions.

10.1.2 Fully automatic AU analysis

The facial expression analysis system capable of recognising both AUs and emotions, is fully automatic. We are not the first to present a fully automated system. There exists at least one fully automated system that uses appearance based techniques, developed by Dr. Marian Bartlett's group at UCSD, and one that uses geometric features, developed by the robotics laboratory at Carnegie Mellon University. The latter only works for subjects that the system has been trained on. To the best of our knowledge we have presented in this thesis the first fully automatic geometric feature based AU analysis system that is capable of dealing with subjects unknown to the system. We believe that this is an important achievement.

We have integrated our fully automatic facial expression analysis system (including emotion recognition), with an artificial portrait painting program called the Painting Fool. With this successful combination of two machine learning/Artificial Intelligence programs we have entered and won the 2007 Machine Intelligence Competition. This, in turn, has resulted in media attention from many online magazines, a British daily newspaper and two BBC radio programs.

One important aspect of having a fully automatic system is that it allows researchers to take their system out of the lab and into the real world. We have done so with our system for the Machine Intelligence Award and related interviews, and have identified in a short period of time a number of issues that need to be addressed which did not seem a problem with our recordings made under lab conditions. Rigid out-of-plane head motion, tracker noise and facial point detection under difficult lighting conditions (particularly shading from overhead-lighting, a common indoors situation) are all things that have to be worked on (see section 10.2 below). On the other hand, other aspects of the system turned out to be extremely robust. For instance, the point tracker was very robust to adverse lighting conditions and the facial point detector performed very well even for people with glasses, beards, or moustaches and still worked fine with about 15 degrees head rotation in any direction.

10.1.3 Temporal dynamics analysis

We have proposed to use a hybrid SVM-HMM classifier with geometric features selected by GentleBoost for the problem of recognising the temporal phases of AUs. We have shown that this method

performs very well. In terms of classification accuracy, the method achieves an average F1-measure of 53.7% for the onset phase, 65.6% for the apex phase and 45.9% for the offset phase. In terms of timing, when the system correctly identifies an AU to be present, it predicts the start of each temporal phase with an error of less than 2 frames on average (1/12-th of a second) and the average temporal phase duration error is less than 4 frames (1/6-th of a second). In terms of F1-measure classification accuracy, it outperforms an implementation with a multiclass gentleSvm by 3% for the onset phase, 7.3% for the apex phase and by 6.6% for the offset phase.

The ability to exactly recognise when a facial action starts, when it reaches the start and end of its peak and when it has returned to neutral opens up a world of possibilities. It allows us to take the analysis of facial expressions to the next level. Durations and differences in timing can now be studied with high precision. We believe that with this method we have paved the way for many real-world applications. The Emotionally Aware Painting Fool, distinguishing posed from spontaneous brow actions and distinguishing posed from spontaneous smiles are but the first examples of this.

Originally, we determined whether an AU was active during an entire video by analysing the output of the AU-activation detection (event detection, see chapter 6). However, we realised that it might be useful to scrutinise the videos in which an AU was predicted to be active by the AU temporal analysis algorithm. Because the SVM-HMM checks whether an AU follows a correct temporal pattern, it can exclude a large number of videos in which an AU was falsely predicted as positive, thereby increasing the precision of our event detection. The danger of course would be that some true positives would incorrectly be dismissed by the temporal analysis algorithm. This fear turned out to be ungrounded, as we achieved a staggering 25.1% precision increase for the event detection, with only a 1.6% decrease in recall. This translates to a F1-measure increase of 17.1%.

10.1.4 Emotive expression recognition

In this thesis we have shown two important facts. Firstly we have shown that it is straightforward to recognise typical displays of emotions using geometric features. Secondly, that a two-step approach in which we first detect AUs and decide based on the AUs which emotion was shown is very well possible too. While there is a significant loss in performance if we employ the two-step approach with automatically detected AUs as input, we have also shown that recognition rates are nearly perfect if the AU detection would be perfect. The added benefits of knowing not only what emotive expression was shown, but also what facial muscles were used to produce this prototypic expression, and what the temporal dynamics of these AUs were make a two-step approach very attractive.

Another very important benefit of the two-step approach is that we can decouple the purely objective facial muscle analysis from the more subjective facial expression analysis. An AU detector can be trained without any knowledge about the intention of the person who made the facial expression. Conversely, an expression analyser, e.g. an emotive expression recognition system, can be trained using AU activation information only. Since AU activation data is very sparse (32 dimensional), creating a huge database of facial expressions is not constricted by disk space anymore. Training systems that can recognise a novel expression or retraining systems when definitions of expressions have changed becomes an easy task. This would allow researchers who are more interested in computer vision and pattern recognition to focus solely on the purely technical problem of AU analysis while other researchers with more interest in human behaviour to focus on the meaning of the shown expressions.

10.1.5 MMI-Facial Expression Database

As part of the work for this thesis, we have created the largest publicly available dataset of FACS-coded videos and static images of facial expressions. The database contains over 3000 videos and high-resolution static images. It is the first to contain recordings of both posed and spontaneous facial expressions, dual-view recordings of facial expressions (frontal and profile), and a very large amount of single-AU activations.

The database already has hundreds of users around the world, with new applications for access to the database reaching us on a daily basis. Our intention is to expand this database so that it can be used to benchmark various facial expression recognition proposals. We intend to create a benchmarking protocol based on this database, with various options to test separate issues such as occlusions, head rotation or lighting effects. A part of the data would be held unavailable to the database users to allow us to organise an AU recognition competition.

10.1.6 Distinguishing posed from spontaneous facial actions

One of the important applications of the automatic AU analysis system proposed in this thesis is to be able to automatically infer what the meaning of a facial expression was. We have shown in chapter 9 that it is indeed possible to use our proposed methods to gain an increased understanding of human (facial) behaviour. In two separate studies, we have used the information about the timing of the temporal phases of AUs to compute a number of high-level features such as the duration of the temporal phases, the velocity with which a facial point moved on average during the onset phase or the number of peaks present in a facial action. In both studies, we have shown that with this type of

information it is possible to distinguish between two different classes of facial behaviour; posed and spontaneous facial actions.

In the first study, we solely used information from the face to distinguish between deliberate (posed) and involuntary (spontaneous) brow actions. This resulted for a dataset of 114 deliberate and 70 spontaneous brow actions in a F1-measure of 87.3%. The second study fused three visual cues to distinguish between deliberate and spontaneous smiles. Fusing the visual cues turned out to greatly improve results, compared to using a single visual cue. A 5.1% increase in F1-measure was obtained when using early fusion. Information from facial point motion, head motion and shoulder motion was combined in a number of different fusion strategies. The best performing fusion strategy for this problem turned out to be late fusion. In this strategy a separate classifier is learned for each visual cue, and a fusion rule aggregates the three classifier results into a single posterior probability. With this approach, a 94.8% F1-measure was achieved.

While the results for both studies are extremely good, we have to keep in mind that the data for the posed and spontaneous classes were recorded for different purposes, under different conditions. While we have tried everything possible to remove database-specific features such as the framerate or the distance of the face to the camera, some caution about the impact of the results is necessary. Still, we believe that it is possible with our proposed system to distinguish between various types of facial behaviour. Future work will require us to test our proposed methods on data that was specifically recorded to evaluate our human behaviour analysis system.

10.2 Future Work

While we have made a number of improvements to the current state of the art in the field of automatic facial expression recognition with the work presented in this thesis, a number of problems remain unsolved and need to be addressed in future work. We are aware that the current version of our system cannot cope with occlusions of (parts of) the face: neither on the level of facial point detection, facial point tracking or the inference of AU morphology/temporal dynamics. This issue has to be addressed on all levels if a system that aims to function under real-world conditions is to be achieved.

Similarly, the current system is not capable of detecting AUs and their temporal dynamics with high accuracy when out-of-image-plane head rotation occurs. To be able to correct for this type of head-pose, we cannot work with the assumption that all facial points lie on a plane anymore. We either need to use a form of 3-dimensional model to accurately detect the head pose and register the facial point tracking data back to a frontal-view representation, or we could divide the head-pose space into a

number of sub-spaces in which the planarity assumption holds locally and train a different AU analysis system for each of the subspaces. Another problem caused by large head rotation is directly linked to the problem of (self) occlusion: when the head is rotated too much with respect to the camera, some facial points simply disappear out of view as they get occluded by other facial features (e.g. the inner eyepoints which get occluded by the nose)¹.

Currently our system assumes that the face initially displays a neutral expression. Of course, in reality we do not know what expression is shown when a user is introduced to the system. Depending on the application, this may or may not be a problem. The two system elements that would need to be changed to be able to deal with non-neutral initial expressions are the facial point detector and the AU analysis system. While we think it will not be too hard to adapt the facial point detector to cope with this issue, it is quite well possible that the facial analysis system will need to analyse a large number of frames before it can be certain about what AUs are activated, and when a neutral expression has began.

Apart from being able to deal with more real-life situations, described above, there are two avenues of future research that we would like to pursue. First of all, we would like to create a system that fuses geometric with appearance based techniques. The two approaches seem to complement each other, and the idea would be that together they can overcome each other's weaknesses (e.g., geometric feature-based approaches are more robust to registration errors while appearance based techniques seem to be more accurate at detecting minute changes in facial actions). The fundamental question for such a hybrid approach would be how to fuse the information from both approaches. If we would fuse at too low a level, the individual strengths of the two approaches might get lost. But if we fuse at too high a level, the two approaches cannot use the information contained in the other channel for their own analysis. Thus, the level of fusion would need to be studied.

The second avenue of future research that we would like to pursue is the recognition of AU intensity. This is an area of research in which to date very little has been done. Yet explicit knowledge about the intensity of facial actions would be of great value to human facial behaviour analysis.

¹Incidentally, this is exactly a proof of the non-planar characteristic of the face. If all facial points would lie on a plane, they would all remain visible until the normal of the plane is perpendicular to the viewing angle of the camera, at which point all facial points on the plane would disappear out of view together

Bibliography

- [Aarts, 2005] Aarts, E. (2005). Ambient intelligence drives open innovation. *ACM Interactions*, 12(4):66–68.
- [Ambadar et al., 2005] Ambadar, Z., Schooler, J., and Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410.
- [Ambady and Rosenthal, 1992] Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 11(2):256–274.
- [Anderson and McOwan, 2006] Anderson, K. and McOwan, P. (2006). A real-time automated system for recognition of human facial expressions. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 36(1):96–105.
- [Andrieu et al., 2004] Andrieu, C., Doucet, A., Singh, S., and Tadic, V. (2004). Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92(3):423–438.
- [Ashraf et al., 2007] Ashraf, A., Lucey, S., Chen, T., Cohn, J., and Prkachin, K. (2007). The painful face - pain expression recognition using active appearance models. In *Int'l conf. multimedia and interfaces*, pages 9–14.
- [Baron-Cohen et al., 2004] Baron-Cohen, S., Golan, O., Wheelwright, S., and Hill, J. (2004). *A New Taxonomy of Human Emotions*. Kingsley Publishers.
- [Bartlett et al., 2003] Bartlett, M., Littlewort, G., Braathen, B., Sejnowski, T., and J.R. Movellan, J. (2003). A prototype for automatic recognition of spontaneous facial actions. *Advances in Neural Information Processing Systems*, 15:1271–1278.

- [Bartlett et al., 2004] Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., and Movellan, J. (2004). Machine learning methods for fully automatic recognition of facial expressions and actions. *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, 1:592–597.
- [Bartlett et al., 1999] Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263.
- [Bartlett et al., 2006] Bartlett, M. S., Littlewort, G., Frank, M. G., Lainscsek, C., Fasel, I., and Movellan, J. (2006). Fully automatic facial action recognition in spontaneous behavior. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 223–230.
- [Bassili, 1978] Bassili, J. N. (1978). Facial motion in the perception of faces and of emotional expression. *J. Experimental Psychology*, 4(3):373–379.
- [Bouvard and Morgan, 1998] Bouvard, H. and Morgan, N. (1998). Hybrid hmm/ann systems for speech recognition: Overview and new research directions. *Lecture Notes in Artificial Intelligence*, pages 389–417.
- [Bradley and Lang, 2000] Bradley, M. and Lang, P. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, 37:204–215.
- [Buller et al., 1994] Buller, D., Burgoon, J., White, C., and Ebesu, A. (1994). Interpersonal deception: Vii. behavioral profiles of falsification, equivocation and concealment. *Journal of Language and Social Psychology*, 13(5):366–395.
- [Burr et al., 1989] Burr, D., Morrone, M., and Spinelli, D. (1989). Evidence for edge and bar detectors in human vision. *Vision Research*, 29(4):419–431.
- [Cacioppo and Berntson, 1994] Cacioppo, J. and Berntson, G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, 115:401–423.
- [Camurri et al., 2005] Camurri, A., Volpe, G., Poli, G. D., and Leman, M. (2005). Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia*, 12(1):43–53.
- [Chang et al., 2006] Chang, Y., Hu, C., Feris, R., and Turk, M. (2006). Manifold based analysis of facial expression. *Image and Vision Computing J.*, 24(6):605–614.
- [Chen et al., 2004] Chen, L., Zhang, L., Zhang, H., and Abdel-Mottaleb, M. (2004). 3d shape constraint for facial feature localization using probabilistic-like output. In *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures*, pages 302–307.

- [Cohen et al., 2003] Cohen, I., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003). Facial expression recognition from video sequences - temporal and static modeling. *Comp. Vision, and Image Understanding*, 91:160–187.
- [Cohen et al., 2002] Cohen, I., Sebe, N., Garg, A., and Huang, T. (2002). Facial expression recognition from video sequences. In *Proc. Int'l conf. Multimedia and Expo*, volume 2, pages 121–124.
- [Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.
- [Cohn and Ekman, 2005] Cohn, J. and Ekman, P. (2005). Measuring facial action by manual coding, facial emg, and automatic facial image analysis. In J. A. Harrigan, R. R. . K. S., editor, *Handbook of nonverbal behavior research methods in the affective sciences*, pages 9–64. Oxford University Press. New York.
- [Cohn et al., 2004a] Cohn, J., Reed, L., Moriyama, T., Xiao, J., Schmidt, K., and Ambadar, Z. (2004a). Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In *Proc. of the Sixth IEEE Int'l Conf. on Automatic Face and Gesture Recognition (FG'04)*, pages 129 – 138.
- [Cohn, 2006] Cohn, J. F. (2006). Foundations of human computing: Facial expression and emotion. *Proc. ACM Int'l Conf. Multimodal Interfaces*, 1:610–616.
- [Cohn et al., 2004b] Cohn, J. F., Reed, J. F., Ambadar, Z., Xiao, J., and Moriyama, T. (2004b). Automatic analysis and recognition of brow actions in spontaneous facial behavior. *Proc. IEEE Int'l Conf. Systems, Man & Cybernetics*, 1:610–616.
- [Cohn and Schmidt, 2004] Cohn, J. F. and Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *J. Wavelets, Multi-resolution and Information Processing*, 2(2):121–132.
- [Cowie et al., 2005] Cowie, R., Douglas-Cowie, E., and Cox, C. (2005). Beyond emotion archetypes: databases for emotion modeling using neural networks. *Neural Networks*, 18:371–388.
- [Cristianini and Shawe-Taylor, 2000] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning*. Cambridge University Press.
- [Cristinacce and Cootes, 2003] Cristinacce, D. and Cootes, T. (2003). Facial feature detection using adaboost with shape constrains. In *British Machine Vision Conference*, pages 231–240.
- [Cristinacce and Cootes, 2006] Cristinacce, D. and Cootes, T. (2006). Facial feature detection and tracking with automatic template selection. *IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pages 429–434.

- [Cunningham et al., 2004] Cunningham, D., Kleiner, M., Wallraven, C., and Bülthoff, H. (2004). The components of conversational facial expressions. *Proc. ACM Int'l Symposium on Applied Perception in Graphics and Visualization*, pages 143–149.
- [Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*.
- [de Bologne, 1862] de Bologne, G. D. (1862). *Mechanisme de la Physionomie*. Translation: The mechanism of human facial expression. Cambridge University Press, New York, 1990.
- [de C. Williams, 2002] de C. Williams, A. C. (2002). Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–488.
- [DePaulo, 2003] DePaulo, B. (2003). Cues to deception. *Psychological Bulletin*, 129(1):74–118.
- [Donato et al., 1999] Donato, G., Bartlett, M., Hager, J., Ekman, P., and Sejnowski, T. (1999). Classifying facial actions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(10):974–989.
- [Douglas-Cowie et al., 2003] Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Commun.*, 40(1-2):33–60.
- [Ekman, 1982] Ekman, P. (1982). *Face of man: Universal expression in a new guinea village*. Garland. New York.
- [Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6:169–200.
- [Ekman, 2003] Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of New York Academy of Sciences*, 1000:105–221.
- [Ekman and Friesen, 1978] Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press. Palo Alto, California, USA.
- [Ekman and Friesen, 1969] Ekman, P. and Friesen, W. V. (1969). The repertoire of of nonverbal behavior: categories, origins, usage, and coding. *Semiotica*, 1:49–98.
- [Ekman and Friesen, 1974] Ekman, P. and Friesen, W. V. (1974). Detecting deception from body or face. *J. Nonverb. Behav.*, 29:288–298.
- [Ekman and Friesen, 1976] Ekman, P. and Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1:56–75.
- [Ekman and Friesen, 1982] Ekman, P. and Friesen, W. V. (1982). Felt, false and miserable smiles. *J. Nonverb. Behav.*, 6(4):238–252.

- [Ekman et al., 2002] Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System*. A Human Face. Salt Lake City.
- [Ekman et al., 1981] Ekman, P., Hager, J., and Friesen, W. V. (1981). The symmetry of emotional and deliberate facial actions. *Psychophysiology*, 18(2):101–106.
- [Ekman and Rosenberg, 2005] Ekman, P. and Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System*. Oxford University Press. Oxford, UK.
- [Essa and Pentland, 1997] Essa, I. and Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763.
- [Faigan, 1990] Faigan, G. (1990). *The Artist's guide to Facial Expressions*. Watson-Guphill Publications.
- [Farkas and Munro, 1987] Farkas, L. and Munro, I. (1987). *Anthropometric facial proportions in medicine*. Thomas. Springfield, Ill., USA.
- [Fasel et al., 2004] Fasel, B., Monay, F., and Gatica-Perez, D. (2004). Latent semantic analysis of facial action codes for automatic facial expression recognition. In *Proc. ACM SIGMM Int'l workshop on Multimedia information retrieval*, pages 181–188.
- [Fasel et al., 2005] Fasel, I., Fortenberry, B., and Movellan, J. (2005). A generative framework for real time object detection and classification. *Comp. Vision, and Image Understanding*, 98(1):181–210.
- [Feris et al., 2002] Feris, R., Gemell, J., Toyama, K., and Krger, V. (2002). Hierarchical wavelet networks for facial feature localization. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 118–123.
- [Field, 1978] Field, D. (1978). Relations between the statistics of natural images and the response properties of cortical cells. *J. Optics Society America A*, 4(2):2379–2394.
- [Flecha-Garcia, 2001] Flecha-Garcia, M. (2001). Facial gestures and communication: what induces raising eyebrow movements in map task dialogues. *Proc. Theoretical and Applied Linguistics - Postgraduate Conf.*
- [Fogel et al., 2006] Fogel, A., Hsu, H., Shapiro, A., Nelson-Goens, G., and Secrist, C. (2006). Effects of normal and perturbed social play on the duration and amplitude of different types of infant smiles. *Developmental Psychology*, 42(3):159–173.

- [Forlizzi, 2005] Forlizzi, J. (2005). Robotic products to assist the aging population. *ACM Interactions Special Issue on Human-Robot Interaction*, 12(2):16–18.
- [Freund and Schapire, 1999] Freund, Y. and Schapire, R. (1999). A short introduction to boosting. *Society for Artificial Intelligence*, 14(5):771–780.
- [Freund and Schapire, 2000] Freund, Y. and Schapire, R. (2000). Discussion of the paper 'additive logistic regression: a statistical view of boosting'. *Annals of Statistics*, 38(2):391–393.
- [Fridlund, 1994] Fridlund, A. (1994). *Human Facial Expression: An Evolutionary View*. Academic Press. San Diego, CA.
- [Friedman et al., 2000] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374.
- [Fumera and Roli, 2005] Fumera, G. and Roli, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(6):942–956.
- [Georghiades et al., 2001] Georghiades, A., Belhumeur, P., and Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):643–660.
- [Gokturk et al., 2002] Gokturk, S. B., Tomasi, C., Girod, B., and Bouguet, J.-Y. (2002). Model-based face tracking for view-independent facial expression recognition. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 287–293.
- [Goleman, 1995] Goleman, D. (1995). *Emotional intelligence: why it can matter more than IQ*. Bantam Books. New York, USA.
- [Graf et al., 2004] Graf, H., Cosatto, E., Bottou, L., Dourdanovic, I., and Vapnik, V. (2004). Parallel support vector machines: The cascade svm. In *Advances in Neural Information Processing Systems*.
- [Gralewski et al., 2006] Gralewski, L., Campbell, N., and Voak, I. (2006). Using a tensor framework for the analysis of facial dynamics. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 217–222.
- [Gross, 2005] Gross, R. (2005). Face databases. In S.Li and A.Jain, editors, *Handbook of Face Recognition*. Springer, New York.
- [Gu and Ji, 2004] Gu, H. and Ji, Q. (2004). An automated face reader for fatigue detection. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 111–116.

- [Gu and Ji, 2005] Gu, H. and Ji, Q. (2005). Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, 16(2):105–115.
- [Gunes and Piccardi, 2007] Gunes, H. and Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345.
- [Guo and Dyer, 2005] Guo, G. and Dyer, C. (2005). Learning from examples in the small sample case - face expression recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(2):105–115.
- [Hayashi and Hasegawa, 2006] Hayashi, S. and Hasegawa, O. (2006). A detection technique for degraded face images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1512.
- [Heller and Haynal, 1994] Heller, M. and Haynal, V. (1994). The faces of suicidal depression (translation), les visages de la depression de suicide. *Cahiers Psychiatriques Genevois (Medecine et Hygiene Editors)*, 16:107–117.
- [Hess et al., 1997] Hess, U., Blairy, S., and Kleck, R. (1997). The intensity of emotional facial expressions and decoding accuracy. *J. Nonverbal Behaviour*, 21(4):241–257.
- [Hess et al., 1989] Hess, U., Kappas, A., McHugo, G., Kleck, R., and Lanzetta, J. (1989). An analysis of the encoding and decoding of spontaneous and posed smiles: the use of facial electromyography. *J. of Nonverbal Behavior*, 13(2):121–137.
- [Hess and Kleck, 1990] Hess, U. and Kleck, R. E. (1990). Differentiating emotion elicited and deliberate emotional facial expressions. *European J. of Social Psychology*, 20(5):369–385.
- [Holden and Owens, 2002] Holden, E. and Owens, R. (2002). Automatic facial point detection. In *Proc. Asian Conf. Computer Vision*, pages 731–736.
- [Hu et al., 2003] Hu, C., Feris, R., and Turk, M. (2003). Real-time view-based face alignment using active wavelet networks. In *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures*, pages 215–221.
- [Hudlicka, 2003] Hudlicka, E. (2003). To feel or not to feel: the role of affect in human-computer interaction. *Int. Journal of Human-Computer Studies*, 59(1–2):1–32.
- [Humphreys et al., 1993] Humphreys, G., Donnelly, N., and Riddoch, M. (1993). Expression is computed separately from facial identity and it is computed separately for moving and static faces - neuropsychological evidence. *Neuropsychologica*, 21:173–181.

- [Isard and Blake, 1998] Isard, M. and Blake, A. (1998). Condensation - conditional density propagation for visual tracking. *Int'l J. of Computer Vision*, 29(1):5–28.
- [J. et al., 1998] J., M., Lyons, Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *AFG*, pages 200–205.
- [Jiang, 2004] Jiang, W. (2004). Boosting with noisy data: Some views from statistical theory. *Neural Computation*, 16(4):789–810.
- [Jones and Palmer, 1987] Jones, J. and Palmer, L. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiology*, 58(6):1233–1258.
- [Kaliouby and Robinson, 2004] Kaliouby, R. E. and Robinson, P. (2004). Real-time inference of complex mental states from facial expressions and head gestures. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 3, page 154.
- [Kalman, 1960] Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Engineering*, 82:35–45.
- [Kanade et al., 2000] Kanade, T., Cohn, J., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 46–53.
- [Kapoor and Picard, 2005] Kapoor, A. and Picard, R. (2005). Multimodal affect recognition in learning environments. In *Proc. of the ACM Int'l Conf. on Multimedia*, pages 677–682.
- [Karpouzis et al., 2007] Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaïou, A., Malatesta, L., and Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal and bodily expressions recognition. *Lecture Notes in Artificial Intelligence*, 4451:92–116.
- [Keltner and Ekman, 2004] Keltner, D. and Ekman, P. (2004). *Handbook of Emotions*. Guilford Press. New York.
- [Kendon, 1973] Kendon, A. (1973). *Comparative Behavior of Primates*. Academic Press. New York.
- [Kirchner et al., 2006] Kirchner, T., Sayette, M., Cohn, J., Moreland, R., and Levine, J. (2006). Effects of alcohol on group formation among male social drinkers. *Journal of Studies on Alcohol*, 67(5):785–794.
- [Kotsia and Pitas, 2007] Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187.

- [Kruger et al., 2005] Kruger, S., Schaffner, M., Katz, M., Andelic, E., and Wendemuth, A. (2005). Speech recognition with support vector machines in a hybrid system. In *Interspeech*, pages 993–996.
- [Kuncheva, 2002] Kuncheva, L. I. (2002). A theoretical study on six classifier fusion strategies. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(2):281–286.
- [Li and Jain, 2005] Li, S. and Jain, A. (2005). *Handbook of Face Recognition*. Springer. New York.
- [Li et al., 2005] Li, S., Lu, X., Hou, X., Peng, X., and Cheng, Q. (2005). Learning multiview face subspaces and facial pose estimation using independent component analysis. *IEEE Transaction on Image Processing*, 14(6):705 – 712.
- [Lien et al., 1998] Lien, J., Kanade, T., Cohn, J., and Li, C. (1998). Subtly different facial expression recognition and expression intensity estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 853–859.
- [Lienhart et al., 2003] Lienhart, R., Kuranov, A., and Pisarevsky, V. (2003). Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Proc. 25th German Pattern Recognition Symposium*, pages 297–304.
- [Littlewort et al., 2004] Littlewort, G., Bartlett, M., Fasel, I., Susskind, J., and Movellan, J. (2004). Dynamics of facial expression extracted automatically from video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 3.
- [Littlewort et al., 2006] Littlewort, G., Bartlett, M., Fasel, I., Susskind, J., and Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Int'l J. on Image and Vision Computing*, 24(6):615–625.
- [Littlewort et al., 2007] Littlewort, G., Bartlett, M., and Lee, K. (2007). Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Int'l conf. on multi-modal interfaces*, pages 15–21.
- [Lucas and Kanade, 1981] Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Int'l Joint Conference on Artificial Intelligence*, pages 674–679.
- [Lucey et al., 2007] Lucey, S., Ashraf, A., and Cohn, J. (2007). Investigating spontaneous facial action recognition through aam representations of the face. In Kurihara, K., editor, *Face Recognition Book*. Pro Literatur Verlag, Mammendorf, Germany.

- [Mandler, 1984] Mandler, G. (1984). *Mind and body: Psychology of emotion and stress*. W.W. Norton and Company. New York.
- [Martinez and Benavente, 1998] Martinez, A. and Benavente, R. (1998). The ar face database. Technical Report 24, Computer Vision Center at UAB.
- [McCowan et al., 2005] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., and Karaiskos, V. (2005). The ami meeting corpus. In Noldus, L. P. J. J., editor, *International conference on methods and techniques in behavioral research; Proceedings of measuring behaviour 2005*, pages 137–140, Wageningen, The Netherlands. Wageningen:.
- [Mehrabian and Russell, 1973] Mehrabian, A. and Russell, J. (1973). A measure of arousal seeking tendency. *Environment and Behavior*, 5:315–333.
- [Merkx et al., 2007] Merkx, P., Truong, K., and Neerincx, M. (2007). Inducing and measuring emotion through a multiplayer first-person shooter computer game. In *Proceedings of Computer Games Workshop*. Amsterdam, The Netherlands.
- [Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.
- [Moghaddam and Pentland, 1997] Moghaddam, B. and Pentland, A. (1997). Probabilistic visual learning for object recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):696–710.
- [Ortony and Turner, 1990] Ortony, A. and Turner, T. (1990). What’s basic about basic emotions? *Psychological review*, 97:315–331.
- [Osadchy et al., 2007] Osadchy, M., Jacobs, D. W., and Lindenbaum, M. (2007). Surface dependent representations for illumination insensitive image comparison. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(1):98–111.
- [O’Toole et al., 2005] O’Toole, A., Harms, J., Snow, S., Hurst, D., Pappas, M., Ayyad, J., and Abdi, H. (2005). A video database of moving faces and people. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):812–816.
- [Pantic and Bartlett, 2007] Pantic, M. and Bartlett, M. (2007). Machine analysis of facial expressions. In Delac, K. and Grgic, M., editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing. Vienna, Austria.
- [Pantic and Patras, 2005] Pantic, M. and Patras, I. (2005). Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Proc. IEEE Int’l Conf. on Systems, Man and Cybernetics*, pages 3358–3363.

- [Pantic and Patras, 2006] Pantic, M. and Patras, I. (2006). Dynamics of facial expressions - recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 36(2):433–449.
- [Pantic et al., 2006] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. S. (2006). Front-end of human computing: Machine analysis of human behavior. *Proc. Int'l Conf. Multimodal Interfaces*.
- [Pantic and Rothkrantz, 2004] Pantic, M. and Rothkrantz, L. (2004). Case-based reasoning for user-profiled recognition of emotions from face images. *Proc. Int'l Conf. Multimedia & Expo*, pages 391–394.
- [Pantic and Rothkrantz, 2000] Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445.
- [Pantic and Rothkrantz, 2003] Pantic, M. and Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE*, 91(9):1370–1390.
- [Pantic et al., 1998] Pantic, M., Rothkrantz, L. J. M., and Koppelaar, H. (1998). Automation of non-verbal communication of facial expressions. *Proc. Conf. Euromedia*, pages 86–93.
- [Pantic et al., 2005a] Pantic, M., Sebe, N., Cohn, J., and Huang, T. (2005a). Affective multimodal human-computer interaction. In *Proc. of the ACM Int'l Conf. on Multimedia*, pages 669–676.
- [Pantic et al., 2005b] Pantic, M., Valstar, M. F., Rademaker, R., and Maat, L. (2005b). Web-based database for facial expression analysis. *Proc. Int'l Conf. Multimedia & Expo*, pages 317–321.
- [Park and Jain, 2006] Park, U. and Jain, A. K. (2006). 3d face reconstruction from stereo video. pages 41–41.
- [Parke, 1974] Parke, F. I. (1974). *A parametric model for human faces*. PhD thesis.
- [Patras and Pantic, 2004] Patras, I. and Pantic, M. (2004). Particle filtering with factorized likelihoods for tracking facial features. *Proc. Int'l Conf. Automatic Face & Gesture Recognition*, pages 97–102.
- [Patras and Pantic, 2005] Patras, I. and Pantic, M. (2005). Tracking deformable motion. *Proc. Int'l Conf. Systems, Man and Cybernetics*, pages 1066–1071.
- [Pentland et al., 1994] Pentland, A., Moghaddam, B., and Starner, T. (1994). View based and modular eigenspaces for face recognition. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 84–91.

- [Pitt and Shephard, 1999] Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filters. *J. Am. Statistical Association*, 94(446):590–616.
- [Platt, 2000] Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola, A., Bartlett, P., Schoelkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT press. Cambridge, MA.
- [Platt and Badler, 1981] Platt, S. M. and Badler, N. I. (1981). Animating facial expressions. In *SIGGRAPH '81: Proc. of the 8th annual conference on Computer graphics and interactive techniques*, pages 245–252, New York, NY, USA. ACM Press.
- [Plutchik, 1980] Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. Harper and Row. New York.
- [Porikli, 2005] Porikli, F. (2005). Integral histogram: A fast way to extract histograms in cartesian spaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 829–836.
- [Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 286.
- [Reinders et al., 1996] Reinders, M., Koch, R., and Gerbrands, J. (1996). Locating facial features in image sequences using neural networks. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 230–235.
- [Roisman et al., 2004] Roisman, G., Tsai, J., and Chiang, K. (2004). The emotional integration of childhood experience: physiological, facial expressive, and self-reported emotional resonance during the adult attachment interview. *Developmental Psychology*, 40(5):776–789.
- [Rosenberg et al., 1998] Rosenberg, E., Ekman, P., and Blumenthal, J. (1998). Facial expression and the affective component of cynical hostility in male coronary heart disease patients. *J. on Health Psychology*, 17(4):376–380.
- [Rosenberg et al., 2001] Rosenberg, E., Ekman, P., Jiang, W., Babyak, M., Coleman, R., Hanson, M., O'Connor, C., Waugh, R., and Blumenthal, J. (2001). Linkages between facial expressions of anger and transient myocardial ischemia in men with coronary artery disease. *Emotion*, 1(2):107–115.
- [Rowley et al., 1998] Rowley, H., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):23–38.

- [Russell, 1980] Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.
- [Russell, 1995] Russell, J. (1995). Facial expression of emotion: What lies beyond minimal universality. *Psychological Bulletin*, 118:379–391.
- [Russell and Fernandez-Dols, 1997] Russell, J. and Fernandez-Dols, J. (1997). *The psychology of facial expression*. Cambridge University Press, New York.
- [Samal and Iyengard, 1992] Samal, A. and Iyengard, P. A. (1992). Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition*, 25(1):65–77.
- [Samaria and Harter, 1994] Samaria, F. and Harter, A. (1994). Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*, pages 138–142.
- [Sayette et al., 1992] Sayette, M., Smith, D., Breiner, M., and Wilson, G. (1992). The effect of alcohol on emotional response to a social stressor. *Journal of Studies on Alcohol*, 53:541–545.
- [Schmidt et al., 2006] Schmidt, K., Ambadar, Z., Cohn, J., and Reed, I. (2006). Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *J. Nonverbal Behavior*, 30(1):37–52.
- [Silva et al., 2005] Silva, P. R. D., Kleinsmith, A., and Bianchi-Berthouze, N. (2005). Towards unsupervised detection of affective body posture nuances. *Int. Conf. Affective Computing and Intelligent Interaction*, pages 32–40.
- [Sim et al., 2003] Sim, T., Baker, S., and Bsat, M. (2003). The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618.
- [Smith and Ellsworth, 1985] Smith, C. and Ellsworth, P. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48:813–838.
- [Smith, 1995] Smith, D. (1995). Natural born liars. *Scientific American*.
- [Sokolova et al., 2006] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021. SpringerLink.
- [Solina et al., 2003] Solina, F., Peer, P., Batagelj, B., Juvan, S., and Kova, J. (2003). Color-based face detection in the "15 seconds of fame" art installation. In *Int'l Conf. on Computer Vision / Computer*

- Graphics Collaboration for Model-based Imaging, Rendering, image Analysis and Graphical special Effects*, pages 38–47.
- [Sung and Poggio, 1998] Sung, K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):39–51.
- [Suwa et al., 1978] Suwa, M., Sugie, N., and Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. *Proc. Int. Joint Conf. Pattern Recognition*, pages 408–410.
- [Tekalp and Ostermann, 2000] Tekalp, A. and Ostermann, J. (2000). Face and 2-d mesh animation in mpeg-4. *Signal Processing: Image Communication*, (15):387–421.
- [Terzopoulos and Waters, 1990] Terzopoulos, D. and Waters, K. (1990). Analysis of facial images using physical and anatomical models. In *Proceedings of the International Conference on Computer Vision*, pages 727–732.
- [Tian et al., 2001] Tian, Y., Kanade, T., and Cohn, J. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):97–115.
- [Tian et al., 2005] Tian, Y. L., Kanade, T., and Cohn, J. F. (2005). *Handbook of Face Recognition*. Springer. New York.
- [Tipping, 2000] Tipping, M. (2000). The relevance vector machine. *Advances in Neural Information Processing Systems*, 12:652–658.
- [Tong et al., 2007] Tong, Y., Liao, W., and Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1683–1699.
- [Tong et al., 2006] Tong, Y., W.Liao, and Ji, Q. (2006). Inferring facial action units with causal relations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1623–1630.
- [Torralba et al., 2004] Torralba, A., Murphy, K., and Freeman, W. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 762–769.
- [Turner and Ortony, 1992] Turner, T. and Ortony, A. (1992). Basic emotions: Can conflicting criteria converge? *Psychological Review*, 99:566–571.

- [Valstar and Pantic, 2006] Valstar, M. F. and Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, page 149.
- [Valstar et al., 2006] Valstar, M. F., Pantic, M., Ambadar, Z., and Cohn, J. F. (2006). Spontaneous vs. posed facial behavior: automatic analysis of brow actions. *Proc. ACM Intl. conf. on Multimodal Interfaces*, pages 162–170.
- [Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag. Berlin.
- [Veldhuizen et al., 2003] Veldhuizen, I., Gaillard, A., and Vries, J. D. (2003). The influence of mental fatigue on facial emg activity during a simulated workday. *Biological Psychology*, 63(1):59–78.
- [Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Robust real-time object detection. *Technical report CRL 200001/01*.
- [Vrij et al., 2000] Vrij, A., Edward, K., Roberts, K., and Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, 24(5):239–263.
- [Wallhoff, 2006] Wallhoff, F. (2006). Facial expressions and emotion database.
- [Weiser, 1991] Weiser, M. (1991). The computer for the twenty-first century. *Scientific American*, 265(3):94–104.
- [Wessels et al., 2005] Wessels, L., Reinders, M., Hart, A., Veenman, C., Dai, H., He, T., and van 't Veer, L. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755–3762.
- [Wiskott et al., 1997] Wiskott, L., Fellous, J., Kruger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779.
- [Xiao et al., 2003] Xiao, J., Moriyama, T., Kanade, T., and Cohn, J. F. (2003). Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Int'l. J. Imaging Systems and Technology*, 13(1):85–94.
- [Yan et al., 2003] Yan, S., Hou, X., Li, S., Zhang, H., and Cheng, Q. (2003). Face alignment using view-based direct appearance models. *Int'l Journal on Imaging Systems and Technology*, 13(1):106–112.
- [Yang et al., 2002] Yang, M., Kriegman, D., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(1):34–58.

- [Yin et al., 2006] Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 211–216.
- [Zeng et al., 2008] Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2008). A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*. In press.
- [Zhang and Ji, 2005] Zhang, Y. and Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):699–714.