Yulia Gizatdinova

Automatic Detection of Face and Facial Features from Images of Neutral and Expressive Faces

ACADEMIC DISSERTATION
To be presented with the permission of the Faculty of Information Sciences of the University of Tampere, for public discussion in Pinni auditorium B1100 on January 16th, 2009, at noon.

Department of Computer Sciences
University of Tampere

Dissertations in Interactive Technology, Number 10
Tampere 2009

ACADEMIC DISSERTATION IN INTERACTIVE TECHNOLOGY

Supervisor:	Professor Veikko Surakka, Ph.D., Department of Computer Sciences, University of Tampere, Finland
Opponent:	Professor Heikki Ailisto, Ph.D., VTT Technical Research Centre of Finland, Finland
Reviewers:	Professor Matti Pietikäinen, Ph.D., Department of Electrical and Information Engineering, University of Oulu, Finland Professor Marcos A. Rodrigues, Ph.D., Materials and Engineering Research Institute, Sheffield Hallam University, UK

Acta Electronica Universitatis Tamperensis 806
ISBN 978-951-44-7591-7 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Dissertations in Interactive Technology, Number 10

Department of Computer Sciences
FIN-33014 University of Tampere
FINLAND

ISBN 978-951-44-7506-1
ISSN 1795-9489

Tampereen yliopistopaino Oy
Tampere 2009

Abstract

The aim of the present dissertation was to develop a framework for automatic and expression-invariant localization of faces and prominent facial landmarks such as eyes, eyebrows, nose, and mouth from static images and real-time videos. For this purpose a local edge-based face representation that remains robust regardless of expressive changes in the face was constructed. Using this facial representation, methods of automatic and expression-invariant face and facial landmark localization were developed during the course of this research work.

The performance of the methods developed was evaluated on several databases of facial expressions coded in terms of prototypical facial displays, like happiness and surprise, and facial muscle activations presented alone or in combinations. In general, the results of testing of the method showed that the constructed local edge-based face representation allowed the face and facial landmarks to be located automatically, robustly, and efficiently from static images and streaming videos displaying facial expressions of varying complexity. The complexity of expressions was presented by a variety of deformations in soft facial tissues, variety of mouth appearances including open and tight mouth, visible teeth and tongue, and self-occlusions.

A further aim of this dissertation was to systematically investigate the effect of single and conjoint muscle activations on the performance of the developed localization methods. The knowledge on the deteriorating effect of certain facial behaviours revealed during the course of this work facilitated further the improvement of the developed methods. Finally, new rectangular performance evaluation measures were introduced in order to evaluate the accuracy of the localization output data. The results showed that the proposed rectangular measures were especially useful in the evaluation of both the location and the size of the eye and mouth localization outputs.

The six publications which follow the dissertation summary demonstrate the successive design, implementation, and testing phases of the method development. Emphasizing simplicity, high speed, and low computation cost of the developed methods, I conclude that they can be utilized in various applications of automatic face analysis. For example, they can be utilized in preliminary localization of facial regions for their subsequent processing in which coarse localization is followed by fine feature detection and analysis.

Acknowledgements

I would like to express my deepest respect and sincere gratitude to my supervisor, Prof. Veikko Surakka. I am grateful for him for being so generous with his time and knowledge while patiently guiding my doctoral research and carefully proofreading all my paper and dissertation drafts. I am highly appreciative of his constant assistance in getting funding for my research and for guiding me to enter the graduate school.

I would like to acknowledge the members of the evaluation committee, Prof. Matti Pietikäinen and Prof. Marcos A. Rodrigues, for their time, effort, and valuable comments in reviewing this dissertation. My sincere thanks go to my friend Oleg Špakov for valuable insights on the prototype implementation issues. I thank other friends of mine, Erno Mäkinen, Jouni Erola, and Tatiana Evreinova for discussions we had about various scientific topics and valuable feedback while commenting some of my papers. I am greatly indebted to my teachers in the past, Dr. Dmitry Shaposhnikov, Dr. Lubov Podladchikova, and Dr. Aleksandr Golovan (A.B. Kogan Research Institute for Neurocybernetics, Rostov State University) for getting me interested in computer vision challenges and teaching me programming skills.

I highly appreciate a remarkable job that the head of our department, Prof. Jyrki Nummenmaa, and the head of the TAUCHI research unit, Prof. Kari-Jouko Räihä, are doing in providing good conditions for the researchers. I also thank helpful administrative and IT support staff members. I am grateful for the financial support from the Tampere Graduate School in Information Science and Engineering (TISE), the University of Tampere, and the Academy of Finland during the course of this work. I am thankful to the director of TISE, Prof. Markku Renfors, and TISE coordinator, Dr. Pertti Koivisto, for making the publication of this dissertation possible. I am grateful for the scholarship from the Center for International Mobility (CIMO). Receiving this scholarship motivated me to move to Finland and start completing the degree.

I am deeply and forever indebted to my parents, Natalia and Faiz, for their love and understanding throughout my entire life. I thank Santtu for his support, encouragement, and patience through these years.

This dissertation is dedicated to my mother, for believing in me.

Tampere, 13th of April, 2008
Yulia Gizatdinova

Contents

1 INTRODUCTION 1

2 FACE AND FACIAL FEATURE DETECTION 5

2.1 Face Detection Mechanisms in the Human Visual Cortex 6

2.2 Face Representation and Facial Features 10

2.3 Approaches to Automatic Face and Facial Feature Detection 13

2.4 Performance Evaluation 37

2.5 Applications 41

2.6 Summary 45

3 FACIAL EXPRESSIONS 51

3.1 Introduction to Facial Expressions 53

3.2 Emotion-Based Expression Categorization 54

3.3 Muscle-Based Expression Categorization 57

3.4 Facial Expression Data 59

3.5 Summary 61

4 A FRAMEWORK FOR FACE AND FACIAL LANDMARK LOCALIZATION 63

4.1 Preprocessing 64

4.2 Edge Detection and Edge Map Construction 66

4.3 Edge Orientation Matching 69

4.4 Structural Correction 72

4.5 Performance Evaluation 75

4.6 Execution Times 78

5 INTRODUCTION TO PUBLICATIONS 79

5.1 Landmark Localization from Images Showing Prototypical Facial Expressions 80

5.2 Landmark Localization from Images Showing Expressions of Happiness and Disgust 81

5.3 Effect of Single and Conjoint AUs on Landmark Localization 83

5.4 Landmark Localization from AU-Coded Facial Images 84

5.5 Fully Automatic Landmark Localization from Expressive Images of High Complexity 85

5.6 Real-Time Localization of Frowning and Smiling Faces Under Head Pose Variation 86

6 DISCUSSION 89

7 CONCLUSIONS 99

8 BIBLIOGRAPHY 101

List of Publications

This dissertation consists of a summary and six original publications as listed below. All the publications are reproduced in the dissertation by permission of the publishers. The publications are presented in chronological order of the research development, which did not always correspond to the chronological order of paper publication due to delays in the publication processes. In the text, the publications are referred to by their corresponding Roman numerals.

I. Gizatdinova Y. and Surakka V. (2006). Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (1), IEEE Computer Society, 135-139. p. 117

II. Guizatdinova I. and Surakka V. (2005). Detection of Facial Landmarks from Neutral, Happy, and Disgust Facial Images. *Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'05)*, UNION Agency: Science Press, 55-62. p. 125

III. Gizatdinova Y. and Surakka V. (2008). Effect of Facial Expressions on Feature-Based Landmark Localization in Static Grey Scale Images. *Proceedings of International Conference on Computer Vision Theory and Applications*, INSTICC, 259-266. p. 135

IV. Gizatdinova Y. and Surakka V. (2007). Automatic Detection of Facial Landmarks from AU-Coded Expressive Facial Images. *Proceedings of International Conference on Image Analysis and Processing (ICIAP'07)*, IEEE Computer Society, 419-424. p. 145

V. Gizatdinova Y. and Surakka V. (2008). Automatic Detection of Facial Landmarks from Expressive Images of High Complexity. *Technical report D-2008-9*, Department of Computer Sciences, University of Tampere. p. 153

VI. Erola J., Gizatdinova Y., and Surakka V. (2008). Automatic Real-Time Localization of Frowning and Smiling Faces Under Head Pose Variations. *Proceedings of International Conference on Signal Processing, Computational Geometry and Artificial Vision (ISCGAV'08)*, WSEAS Press, 22-27. (The extended version of this paper was published in *WSEAS Transactions on Signal Processing*, 1 (4), 463-473, 2008.) p. 181

Author’s Research Contributions

The author’s research contributions included the original design of the edge-based method of automatic and expression-invariant face and facial landmark localization from static facial images, iterative development of software prototypes of the designed method, evaluation of the prototypes, analysis of the test results obtained, and writing all publications.

Publication VI was written by the author in collaboration with Jouni Erola, who is the first author of that paper. Publication VI was based on Erola’s MSc dissertation (Erola, 2008) in which the method proposed by the present author was extended for the purpose of real-time face localization under expression and head pose variations. The contribution of the present author was in consulting for the process of the method development, analysing the results of the method testing, and writing the paper.

List of Figures and Tables

Figures

2.1 Simplified visual pathway of face detection in the cerebral cortical areas of the brain (sectional view). The data processing flow goes through areas 1-2-3-4-5-6. p. 8

2.2 Classification of face and facial feature detection methods. p. 14

2.3 Examples of colour-based face detection: (a, c) original colour images and (b, d) the detection result in which bright regions show the detected skin-like areas of the image. p. 16

2.4 The mouth region is detected using the redness property of the segmented regions. Eyes, as areas of high intensity variance in the facial region, are searched above the detected mouth region using heuristics rules. p. 17

2.5 The left image shows a smoothed original facial image. The right image demonstrates the edge orientation template of the face constructed on the basis of the edge map of the image. p. 19

2.6 A Gaborface representation of the facial image demonstrates the convolution results of the image with three Gabor kernels. p. 21

2.7 From left to right: “mean eye” pattern; its wavelet decomposition; “mean eye” pattern at higher resolution; and its wavelet decomposition. The selected features of the two eye patterns are shown as red contours. High intensities correspond to strong edges and low intensities indicate uniform regions (Campadelli *et al.*, 2007). p. 21

2.8 The process of pixel labelling in the image neighbourhood of 3 by 3 pixels to calculate a local LBP code at this pixel position. p. 23

2.9 The left image shows the colour-segmented facial region. The right graph shows the y-axis grey level relief of the whole facial region. The bottom graph shows the x-axis grey level relief of the eye regions. p. 24

2.10 Examples of nose processing (Campadelli *et al.*, 2007). The black horizontal line indicates the base of the nose, the black dots along the nose are the points of maximal symmetry along each row, the black vertical line approximates those points, and the black rectangular marker indicates the tip of the nose. p. 24

2.11 An example of infrared-based eye detection. The left image shows the detected eye pupils (filled circles) and feature points (unfilled circles). The right image shows the verified feature locations. p. 25

2.12 Facial ratio template based on relative brightness of 16 neighbouring regions (grey boxes) and 23 relations between regions (arrows) (Sinha, 1994). p. 28

2.13 The global facial template consists of dark areas which correspond to the prominent facial landmarks. It can be seen that the expression of surprise leads to inconsistency between the template and the locations of the eyebrows and mouth. p. 28

2.14 The left image shows the results of snake-based face contour detection. Reprinted with permission. The top-right image shows snake-based eye detection (Campadelli *et al.*, 2007). From left to right and top to down: eye subimage; edge map; binarized map; initial template position; final template position; and detected eye points. The bottom-right image shows snake-based mouth detection. p. 29

2.15 Face and feature detection by active appearance models (AAM). From left to right and top to bottom: original facial image; segmented face and feature contours; resulting AAM mesh; initial AAM model; AAM model after 2 iterations; and converged AAM model after 12 iterations. p. 30

2.16 Examples of eigenfaces. p. 32

2.17 Coarse-to-fine eye detection (Campadelli *et al.*, 2007). The top row shows low-level eye region detection. The bottom row shows high-level eye position refinement. p. 35

2.18 The left image shows examples of rectangle features. The sum of the pixels which p. 36

lie within the light rectangles is subtracted from the sum of pixels in the dark rectangles. The right image shows rectangle features superimposed on a typical training face. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose.

2.19 Example of the FFFD detection output data in a form of rectangular boxes placed over the found facial regions (Wilson & Fernandez 2006). p. 38

2.20 The detected area for manually annotated face (solid line) is R1 and the detected area for automatically detected face (dashed line) is R2. p. 40

2.21 Multiple application fields of automatic face and feature detection (AFFD). p. 41

3.1 Prototypical facial expressions of emotions proposed by Ekman (1989) and primary directional cues of muscle displacements based on the suggestions of Bassili (1979). p. 55

3.2 Prototypic facial expressions of emotions and corresponding AU codes. p. 58

4.1 Block structure diagram of the major components of the framework for face and facial landmark localization. p. 63

4.2 Image of happiness. Original high-resolution image (*l*=1) on the left and smoothed low-resolution image (*l*=2) on the right. The number of small details is noticeably reduced in the low-resolution image. Image labels are masked by white boxes. p. 65

4.3 Orientation template for detection of local oriented edges, $\varphi = 22.5°$, $i = 0 \div 15$. Orientations $2 \div 6$ and $10 \div 14$ are used for orientation matching. p. 67

4.4 Edge detection and edge map construction: (a) original image of happiness; (b) detected local oriented edges (black regions); and (c) grouped edges after contrast thresholding. p. 69

4.5 Examples of landmark orientation portraits averaged over all datasets. The error bars show plus/minus one standard deviation from the mean values. Reprinted with permission. p. 70

4.6 Examples of individual orientation portraits of facial landmarks from images of Cohn-Kanade facial expression database. p. 71

4.7 From left to right: original image of happiness; edge regions accepted (black rectangles) and discarded (grey areas) by the edge orientation model; final edge map after edge orientation matching. p. 72

4.8 Top row shows landmarks erroneously grouped into one region. Middle row shows landmarks separated by the procedure of edge projection. Bottom row demonstrates the final detection result. p. 73

4.9 Edge maps of facial landmarks: (a) eye regions wrongly detected as one region; (b) eye regions separated by edge projection; (c) nose and mouth wrongly detected as one region; and (d) nose and mouth separated by edge projection. X- and Y-axes show the corresponding pixels in the image. Black dots represent a number of projected edge points per column or row of the upper or lower face landmark candidates respectively. Areas marked with upward diagonal lines show regions of the edge map where edges were eliminated. p. 73

4.10 Examples of the final localization results with “primary” landmarks (rectangles) and “secondary” landmarks (ovals) of the landmark localization. Images are from the Cohn-Kanade AU-Coded Facial Expression Database (Kanade *et al.*, 2000). p. 75

4.11 A set of characteristic points selected to define landmark location in the image: (a) eyebrow and eye; (b) eye region; and (d) mouth. The bounding box containing a landmark was defined by its top-left (*tl*) and bottom-right (*br*) bounding coordinates. The centre point of the landmark was defined by the centre point of the bounding box. p. 77

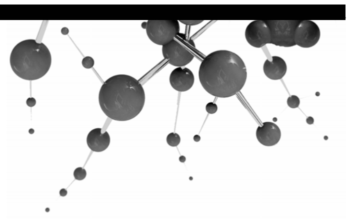
5.1 The research process in chronological order denotes the design, implementation, and evaluation phases of the development of the face and facial landmark localization framework. p. 79

Tables

I Upper and lower face AUs descriptors from the FACS manual (Ekman et al. 2002). p. 57

List of Abbreviations and Acronyms

AAM	Active Appearance Model	p. 30
AAEF	Automatic Analysis of Expressive Faces	p. 51
AdaBoost	Adaptive Boosting Classifiers	p. 36
AFFD	Automatic Face and Feature Detection	p. 41
AFFEX	System for Identifying Affect Expressions	p. 58
AU	Action Unit	p. 57
BMV	Behavioural Model of Vision	p. 26
Cohn-Kanade	The Cohn-Kanade AU-Coded Facial Expression Database	p. 59
EMFACS	Emotional Facial Action Coding System	p. 58
FACS	Facial Action Coding System	p. 57
FACSAID	FACS Affect Interpretation Dictionary	p. 58
FFA	Fusiform Face Area	p. 9
FPS	Frames per Second	p. 40
HCI	Human-Computer Interaction	p. 1
HTI	Human-Technology Interaction	p. 1
JAFFE	Japanese Female Facial Expression Database	p. 59
LBP	Local Binary Patterns	p. 22
MAX	Maximally Discriminative Facial Movement Coding System	p. 58
MPEG-4: SNHC	Object-Based Multi-Media Compression Standard: Synthetic Natural Hybrid Coding	p. 58
NN	Neural Networks	p. 33
POFA	Pictures of Facial Affect Database	p. 59
PCA	Principal Component Analysis	p. 32
SNoW	Sparse Network of Winnow	p. 34
SPI	Seconds per Image	p. 78
SVM	Support Vector Machines	p. 35
V1	Primary Visual Cortex	p. 7
V2	Secondary Visual Cortex	p. 8



1 Introduction

Most human natural interaction takes place through face-to-face communication and it is evident that the face is one of the most important media in human-human interaction. The obvious importance of facial stimuli for humans naturally motivates the idea of also utilizing facial information in human-computer interaction (HCI). The concept of HCI was recently extended into a more general concept of human-technology interaction (HTI). In the HTI concept, the user is actively observed by computers and interacts with them and other computational devices embedded in the environment.

With such a background, it is required that facial information is automatically captured, analysed, and further processed in order to make HTI more intuitive, natural, and intelligent. As postulated by many researchers (Picard, 1997; 2002; Surakka & Vanhala, 2008; Jaimes & Sebe, 2007), facial expressions, head and eye movements, head pose, and gaze direction provide a rich source of information about the user. Considering a face as an input to the computer, different information can be extracted from it and utilized in HTI. For example, the user's identity, age, gender, ethnicity, affective state, focus of attention, and current activity – all this information can be automatically acquired from the user's face. Using this information, the computer can be controlled by or adjusted to the user's needs.

The field of computer vision is a scientific discipline that deals with the technology of building systems of artificial vision. In vision-based HTI, computer vision capabilities can be especially helpful for capturing important visual cues about the user. Generally speaking, computer vision covers areas like detection, recognition, and tracking of visible objects. When applied to automatic face analysis, it includes all tasks related to computer-mediated processing and analysis of visual information contained in the face. Such tasks of automatic face analysis as face and facial feature detection and tracking, head pose estimation and tracking, gaze and eye detection and tracking, lip reading, face recognition, facial expression analysis, gender recognition, ethnicity recognition, age recognition, and action recognition - are some of the many examples of computer vision challenges in HTI. Considerable progress has been achieved and different automatic face analysis schemes have been introduced in the literature. The current progress in computer vision technology enhanced by recent advances in image processing, pattern recognition, and hardware developments makes automatic face analysis possible and applicable in HTI. Computer vision capabilities have already started to advance the situation in HTI to be applicable, for example, to mobile phones (*e.g.* FaceTracker™ [1] and Face Sensing Engine [2]) and gaming technology (*e.g.* EyeToy[3] and Kick Ass Kung-Fu[4]). However, more work is still needed to improve all stages of automatic face analysis.

In this dissertation, one aspect of automatic face analysis, namely, face and feature detection, was addressed. In fact, face and feature detection is the necessary first step in any system of automatic face analysis. Indeed, before any face processing can take place, the facial region(s) should first be detected. Different techniques are applied to static image or streaming video in order to find face-like regions and deliver them as inputs to various systems of automatic face analysis. Automatic face and feature detection can be useful in improving interaction with computers by facilitating the development of perceptive vision-based interfaces which detect the presence and number of users, teleconferencing systems which automatically provide additional bandwidth if a new participant's face is detected, systems of image or video content retrieval and coding, and systems of recognition of the user's facial identity and expressions. Automatic detection of a face and its parts is also useful in many other areas such as security and surveillance improvement in diverse contexts of information society, facial expression analysis, biometrics, education, interactive entertainment, intelligent kiosks, and medical diagnostics.

The considerable potential for applications makes face and feature detection an interesting and challenging area of research. As will be demonstrated later, the goodness of the detection output has an apparent influence on the performance of all following steps of the automatic face analysis. Thus nearly all listed face analysis applications require a face and feature detection scheme that is fully automatic, real-time, robust, and accurate. Over the years, the research in this area has combined great

[1] FaceTracker™, © 2008 FotoNation Inc. at www.fotonation.com, (last accessed July 21, 2008).

[2] Face Sensing Engine (FSE), © 2008 Oki Electric Industry Co., Ltd. at www.oki.com, (last accessed July 21, 2008).

[3] EyeToy, © 2005 Sony at www.eyetoy.com, (last accessed July 15, 2008).

[4] Kick Ass Kung-Fu, © 2003-2005 Animaatiokone Industries at www.animaatiokone.net, (last accessed July 15, 2008).

efforts from researchers in computer vision, pattern recognition, and image processing and analysis. Yet, the development of a face and feature detection system which would satisfy all the requirements is still a challenge.

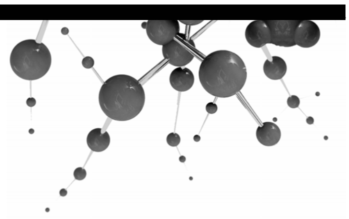
The difficulty in developing face and feature detection systems results from the fact that facial appearance depends significantly on varying environmental conditions and inner properties of the face. The environmental conditions include illumination change, out- and in-plane head rotations, scene complexity, resolution, scale or image size, occlusions, and presence of structural components like eye-glasses. The inner characteristics of the face are those related to ethnicity or skin colour, gender, facial expressions, age-specific facial deformations, and facial hair. Facial expressions have been demonstrated to be one of the most problematic factors which affect the performance of face and feature detection systems (Yacoob *et al*., 1995; Pantic & Rothkrantz, 2000*a*). The detection of mouth and eyes is especially difficult, since these facial features are highly deformable and may vary in shape and colour. In addition, the eyes and mouth are often subject to self-occlusion during expressive reactions.

The current work aimed at contributing to the research in this field by utilizing computer vision capabilities for the task of automatic expression-invariant face and facial feature detection. When I started this research, most of the past research work in this area had focused mainly on the processing of expressionless faces, thus, leaving out the research question of expression invariance. It became apparent that in order to develop an effective expression-invariant face and feature detection system, a new representation of the face was needed. This face representation should remain robust with respect to a variety of facial expressions. Thus, the first research question faced was *"What kind of facial representation is invariant with respect to expressive changes in the face*?" On the other hand, this face representation should describe a face and facial features in a detailed but compact and easy-to-compute way, as the requirement of real-time processing also puts challenges on face and feature detection algorithm implementation and optimization. Further, the earlier studies addressing the issue of expression-invariant face and feature detection have not always provided a careful and systematic approach to report their findings. This posed another important research question as to *"What expressive changes in facial appearance make effect on the performance of face and facial feature detection?"* Another important research question related to the accuracy of face and feature output result was *"What can be considered a correct detection result and what is the allowable rate of detection error?"* So far in the literature only few studies have seriously addressed these issues and much more effort was still needed to set up convenient performance evaluation measures for face and feature detection systems. In addition, the compatibility of face and feature detection output data with requirements for the input data posed by systems of automatic face analysis has to be considered.

Therefore, the main objectives of this dissertation were as follows.

- To find a face representation that remains invariant regardless of deformations in the face brought about by facial muscle activations
- To design methods of expression-invariant face and feature detection in the presence of complex facial expressions from static images and streaming video
- To study the effect of different expressive facial behaviours on the performance of the methods developed for face and facial feature detection
- To develop and test measures for the performance evaluation of the detection methods developed
- To conduct carefully controlled empirical studies in order to test the methods developed on large databases of expressive images and video
- To develop and implement a software framework for face and facial feature detection which combines the face and feature detection methods developed

The dissertation consists of a summary and the six original publications listed on page VI. The summary describes both theoretical and practical aspects of face and feature detection in general and the specific solution to automatic and expression-invariant face and feature detection proposed in this doctoral work. Chapter 2 presents a general review of the existing face and feature detection approaches, techniques, and other related work. Because the main focus was on expression-invariant face and feature detection, recent work is reviewed in this direction. Further, a closer look is taken at three main application fields of face and feature detection which are face recognition, facial expression analysis, and vision-based perceptual HTI. Chapter 3 gives an overview of facial expressions, approaches to facial expression categorization, and a description of the facial expression databases used in this research. The proposed face and feature detection framework is presented in Chapter 4. In Chapter 5, each of the separate publications is related to the context of this dissertation. The experimental results are also presented and discussed in order to demonstrate the course of the research and show the robustness of the face and feature detection methods developed under facial expression variations. Chapter 6 presents a general discussion of the results and possible directions for the future research. Chapter 7 concludes the dissertation.



2 Face and Facial Feature Detection

During the course of this work, it became apparent that among research papers and literature surveys on the topic, the problem of automatic finding of face and facial parts was referred to by different terms, which were sometimes interchangeable. Thus, this chapter starts from a classification of the terminology used in the dissertation and a clarification of the scope of the dissertation in this area.

Generally, face detection can be considered as a specific case of object-class detection in which the task is to find the locations and sizes of all objects belonging to a certain class. Given an image or video frame, *face detection* is defined as a process of automatic finding of the true location and size of the face(s) if there are any, or declaring about their absence. *Face localization* is a simplified detection problem with the assumption that a face is shown in the image or video sequence. Given an image or video frame with a known number of faces (*i.e.* usually one), the task is to find the true face locations and sizes. *Face tracking* refers to the task of face detection or localization as applied specifically to video input data. After a face is detected from the first video frame, the process of face detection in the following frames is guided by information on the face location in the previous frame(s). One of the main requirements for face tracking algorithms is their speed – face tracking has to be fast enough for use in real-time applications. Usually it is assumed that there is only one face in the image or video frame. In the opposite case, the task is called *multiple face detection or tracking*. *Face segmentation* typically refers to a separation of the facial region from the background.

Facial feature detection, localization, and tracking - are all defined in a way similar to that described above for face detection definitions. As will be discussed further, the main features to be detected in the face are usually prominent facial landmarks such as eyes, eyebrows, nose, and mouth. In this respect, the detection and localization of the prominent facial landmarks has recently been referred to as a *landmarking* process (Salah *et al.*, 2007). *Facial feature extraction* generally refers to the process of facial feature search in a broad sense, making no particular distinction between the tasks of feature detection, localization, or tracking. *Face and facial feature detection, localization, and tracking in 3D* are based on 3D face representation that can be retrieved from 2D face representations (*e.g.* multiple facial views received from different cameras).

Due to the nature of the image data used in this dissertation, the main efforts were concentrated on the task of face and facial landmark localization from static facial images and streaming video. Thus, face and feature tracking as well as 3D facial image analysis were excluded from the scope of this dissertation. Later, in Chapters 4 and 5 and the publications following this summary, the task of face and facial feature detection, unless otherwise specified, is referred to as a localization process. In the section that follows, the known biological mechanisms which facilitate face processing in the human visual cortex are described. The importance of this section is in the description of the low-vision pre-attentive biological mechanisms of face detection which were simulated to some extent in this dissertation. A general concept of face representation from the computer vision point of view is given in the next section. Further, a survey is presented reviewing approaches to face and feature detection covering the most significant achievements in this field. Issues related to the performance evaluation of face and feature detectors are moreover discussed. The survey is not intended to provide an exhaustive review of the past work on each of the problems related to automatic face and feature detection, but rather attempts to highlight the most significant aspects of the field. At the end of the chapter, three main target application areas of face and feature detection data are introduced. The application areas are perceptual vision-based HTI, face recognition, and facial expression analysis. The chapter ends with a summary.

2.1 FACE DETECTION MECHANISMS IN THE HUMAN VISUAL CORTEX

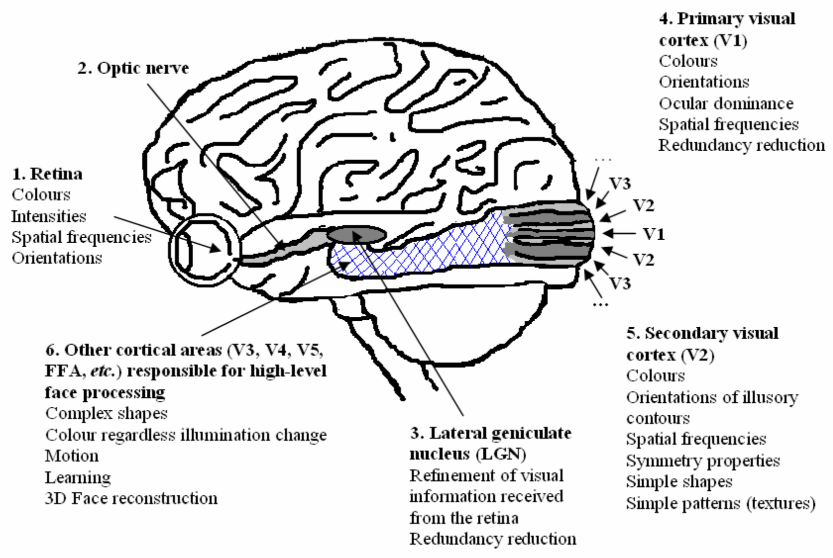
Normally, humans demonstrate an astonishing ability to detect faces in an effortless, fast, and accurate manner from complex natural scenes, even if only limited facial information is available (Sinha *et al.*, 2006). Many studies on functional imaging, neuropsychology, and electrophysiology have attempted to shed light on the question of how face detection is accomplished in the human brain. There is so far no consensus view among different authors on this question. Some authors indicate specific brain areas corresponding to face detection processes (Kanwisher *et al.*, 1997; Gelder & Rouw, 2001; Dailey *et al.*, 2002). Other studies argue that multiple distributed cortical areas of the brain give strong responses when a facial stimulus is presented (Haxby *et al.*, 2001). In spite of some unresolved theoretical implications of these findings, the general consensus is that humans effortlessly detect and process faces due to the effective mechanisms of visual attention (Hubel, 1995; Walther *et al.*, 2002). According to this theory, the human brain simultaneously performs two tasks of face processing - bottom-up and top-down face processing. *Bottom-up face processing* enables mechanisms of low-level pre-attentive vision to select characteristic areas of the image according to various visual stimuli such as intensity, colour, orientation, shape, texture, and others. *Top-down face processing* corresponds to attentive (*i.e.* perceptive) mechanisms of high-level information processing such as understanding, memorizing, and recognition.

It has been suggested (Bruce & Young, 1986) that face detection is the initial stage of any visual face processing in the brain. Early studies have demonstrated that distinguishing between face and non-face objects may occur independently of other face processing functions (Ellis, 1981). These studies demonstrated that face detection may perform successfully while face recognition and facial expression recognition may be severely compromised (Gessler *et al.*, 1989; Heimberg *et al.*, 1992). It has been also shown (Bruce & Humphreys, 1994) that in face detection humans mainly rely on such properties of the visual signal as edges rather than, for example, colours, textures, and shadings. This is confirmed by the fact that humans demonstrate high rates of face detection from both high-quality colour photographs and simple line drawings (Chen *et al.*, 2008), in contrast to face or facial expression recognition. More complex image properties such as patterns and shapes are also extracted from the image, presumably to reconstruct various facial models on the proceeding stages of high-level face processing.

The early processing of a facial signal is performed already in the retina, which is located in the inner layer of the eye. The retina consists of multiple layers of cells. The receptor cells, *rods and cons*, are located at the bottom of the retina. These receptors are sensitive to light. Millions of receptors are packed into the retina. These cells are responsive to the simple properties of the visual signal, such as intensities, edges, and colours. The visual signal is then transferred through the optic nerve to the lateral *geniculate nucleus*, in which the reconstruction of the visual information from the receptive fields[5] of the cells of the retina starts. Then the signal is projected into the primary visual cortex called V1. The V1

[5] The receptive field of a visual neuron is a region of retina in which the presence of light alters the firing of that neuron.

neurons have the ability to respond to somewhat more complex properties of the facial image such as colours, orientations of line segments, and spatial frequencies. V1 next sends the signal to the secondary visual cortex called V2. Although the majority of neurons in V2 are responsive to the same visual properties as the neurons in V1, there are also other neurons which are responsive to complex properties of the image such as simple shapes and patterns. Figure 2.1 demonstrates a simplified visual pathway of face detection in the visual cortical areas of the brain.



4. Primary visual cortex (V1)
Colours
Orientations
Ocular dominance
Spatial frequencies
Redundancy reduction
2. Optic nerve
1. Retina
Colours
Intensities
Spatial frequencies
Orientations
V3
V2
V1
V2
V3
6. Other cortical areas (V3, V4, V5, FFA, *etc.*) responsible for high-level face processing
Complex shapes
Colour regardless illumination change
Motion
Learning
3D Face reconstruction
3. Lateral geniculate nucleus (LGN)
Refinement of visual information received from the retina
Redundancy reduction
5. Secondary visual cortex (V2)
Colours
Orientations of illusory contours
Spatial frequencies
Symmetry properties
Simple shapes
Simple patterns (textures)

Figure 2.1. Simplified visual pathway of face detection in the cerebral cortical areas of the brain (sectional view). The data processing flow goes through areas 1-2-3-4-5-6. Modified from (Ban *et al.*, 2004). Reprinted with permission.[6]

Most cells in the retina, V1, and V2 have a remarkable property of orientation selectivity and respond best to edges at some particular orientation. This property enables the detection of local oriented edges and the definition of their orientations. According to the concept of columnar organization (Hubel & Wiesel [7], 1962; 1974; Orban, 1984; Hubel, 1995), the neighbouring neurons in V1 and V2 have similar orientation selectivity. Together they form an orientation column or iso-orientation domain. A set of orientation columns with a common receptive field forms a module of the cortex called a hypercolumn. The neurons in the hypercolumn affect each other. For example, there is the phenomenon of lateral inhibition, in which there is a reduction of activity in one neuron caused by activity in a neighbouring neuron. Lateral inhibition is useful,

[6] Reprinted from *Neurocomputing*, *56*, 2004, Ban S.-W., Lee M., and Yang H.-S, “A Face Detection Using Biologically Motivated Bottom-Up Saliency Map Model and Top-Down Perception Model”, Fig. 1 (p. 476), copyright © 2003 Elsevier B.V., with permission from Elsevier B.V.

[7] In 1981 David Hubel and Torsten Wiesel were awarded the Nobel Prize in Physiology or Medicine for their discoveries concerning information processing in the biological visual systems.

because it increases the contrast at the edges of objects, thus making it easier to identify the border line between one object and another.

V1 and V2 are surrounded by many other areas of visual cortex. There are feedforward and feedback connections between primary, secondary, and high-level cortical areas. The connections are stronger between the neighbouring areas and weaker between the remote ones. Altogether, they form two major cortical pathways of visual information processing. The *ventral pathway* is thought to be involved in the perception, representation, and recognition of objects by processing their characteristic visual properties, such as shape and colour. It is also associated with long-term memory storage. It begins with V1, goes through V2, then through a cortical area called V4, proceeds to the inferior temporal cortex, and enters the high-level cortical areas associated with cognition and memory. Increasing evidence has been reported about a *fusiform face area* (FFA) in the ventral pathway which is involved in multiple tasks of face processing (Kanwisher *et al.*, 1997; Gelder & Rouw, 2001; Dailey *et al.*, 2002). The FFA unit was postulated to be involved in face detection (Tong *et al.*, 2000), structural encoding of faces (Zion-Golumbic & Bentin, 2007; Chen *et al.*, 2008) and subordinate level categorization of non-face objects (Gauthier *et al.*, 1997). The *dorsal pathway* begins with V1, goes through V2, then to the dorsomedial area, the cortical area called V5, and to the posterior parietal cortex. The dorsal pathway is associated with visual-motor control of the eyes and arms, representation of object size, location, and spatial orientation. Low-vision pre-attentive face processing occurs mainly in the retina, V1 and to some extent in V2. They mainly serve the purpose of perception of primitive features of the visual signal and reduction of the information redundancy from the signal. This property is of great importance because irrelevant information can already be discarded at the early stages of visual signal processing. This allows the high-level processing mechanisms to concentrate on important objects of the visual scene.

With continuous progress in functional imaging, neuropsychology, and electrophysiology, a better understanding of mechanisms of biological vision is acquired yielding options to simulate them in the systems of automatic vision. On account of the astonishing ability of the human brain to achieve nearly perfect face processing results in varying conditions, computer scientists have long endeavoured to develop biologically plausible models of computer vision (Reisfeld, 1993; Herpers *et al.*, 1995; Riesenhuber & Poggio, 1999; Ullman *et al.*, 2002; Walther *et al.*, 2002; Ban *et al.*, 2004; Ban & Lee, 2005; Serre *et al.*, 2005; Shevtsova *et al.*, 2007). As will be demonstrated later in Chapter 4, the analysis of the edge structures utilized in this dissertation has an apparent relevance to the systems of low-level pre-attentive biological vision (Rybak *et al.*, 1990; 2005). As in earlier works (Golovan *et al.*, 2000; 2001; Shaposhnikov *et al.*, 2002), in this dissertation, too, the edge detection module for face and facial landmark localization imitates the hypercolumn neurons of the visual cortex, which are sensitive to different orientations of local edges.

2.2 Face Representation and Facial Features

Let us now consider the general concept of face representation from the perspective of computer vision. Human faces constitute a class of visually similar objects with a rigid structure that does not vary significantly from person to person (*e.g.* typically there are two eyes and the nose is located between eyes and mouth). This makes the task of distinguishing faces from objects of other classes relatively easy. However, within the face class there exist variations which make the detection of facial features a difficult task. Facial appearance varies noticeably with changes in environmental conditions and between and within individuals. Changes in the environmental conditions which typically impair a performance of face and feature detectors are due to illumination, out- and in-plane head rotations, scene complexity, camera characteristics (i.e. image resolution, viewing distance, camera sensor noise), scale or image size, structural components like eye-glasses, and occlusions. The inner properties of the face determine between-individual variations which are due to race and gender, and within-subject variations which are due to facial expressions and age-specific face modifications. Considering such a variety of facial appearances, "The goal of feature extraction is to process the raw pixel data such that variations between objects of the same class (within-class variations) are reduced while variations relevant for separating between objects of different classes (between-class variations) are kept" (Heisele *et al.*, 2000*a*).

Different features can be detected from the facial image, for example, colours, points of interest, line segments and their intersections, contours, grey-scale intensities, wavelet decompositions, first and second derivatives of grey-scale pixel values, and other statistics. These abstract image features are used to represent more general facial features such as facial landmarks - eyes, eyebrows, nose, mouth, chin, cheeks, forehead, hair-line, bridge of the nose, eye pupils, nostrils, face outline, and others. There are two commonly used types of *spatial face representation* - global (*i.e.* holistic) and local face representations. *Global representation* is usually defined as an image feature vector that contains information about the whole facial pattern. Because global face representation considers a face as a whole, it already contains the knowledge of the spatial arrangement of independent facial features. *Local representation* considers distinctive properties of the local regions and points of the face which are usually located in the neighbourhood of the facial landmarks.

A question is what facial features to detect in the image? In the literature, there is so far no general consensus as to which facial features are the best and most frequently used in face and feature detection. Usually the choice of facial features for detection depends on the characteristics of the image and the application of the face and feature detection output data. The most often cited facial features are consistent with those features which seem to be naturally distinctive for humans. These features are called *prominent facial landmarks* like eyes, eyebrows, nose, and mouth (Ekman & Friesen, 1978; Sinha *et al.*, 2006). The prominent landmarks are frequently used in face recognition as they encode critical information about the facial structure. They are also widely used in facial expression recognition because the information about an expression is mainly concentrated around prominent facial landmarks. In addition, prominent facial landmarks are relatively easy to detect as they contain rich edge structures and exhibit more contrast than, for example, cheeks and chin, which contain only few edge structures. The other features are considered to be *secondary facial landmarks*.

Eyes and eye regions. It has been widely acknowledged that eyes are the main prominent facial features for humans followed by mouth, nose, and eyebrows (Ekman & Friesen, 1978; Fasel *et al.* 2005; Sinha *et al.*, 2006). The eye region and eyes convey crucial information about the affective state of the person and the person's facial identity. The direction of the eye gaze shows the focus of attention and the possible intentions of the person. Therefore a robust and non-intrusive automatic eye analysis is crucial for many computer-mediated applications. In HTI, for example, information derived from the eye regions helps in building perceptive user interfaces and user affective state understanding. Changes in the appearance of the eye regions in the face can be successfully used to perform facial expression analysis (Pantic & Rothkrantz, 2000*a,b*) and identify and classify particular affective states of the user (Heishman *et al.*, 2004).

Systems of face identification and verification also rely heavily on the eye regions which provide important biometrical characteristics of a person to be recognized (Wiskott *et al.*, 1997; Petrushan *et al.*, 2005; Wang *et al.*, 2005; Campadelli *et al.*, 2007). Normalization and head pose estimation is done on the basis of eye locations (Burl & Perona, 1996; Gao *et al.*, 2007). Various eye tracking systems use eye pupil positions as input for gaze control applications like typing and pointing in visual graphical user interfaces (Majaranta & Räihä, 2007; Surakka *et al.*, 2004) and as a means of studying pupillary responses to different types of stimuli (Partala and Surakka, 2003). Additionally, eye closure and blinking rate calculation have been utilized in systems of driver fatigue monitoring[8] in which eye closure is

[8] Eye Alert® Fatigue Warning System, © 2005 Highway Safety Group at www.eyealert.com, (last accessed on July 16, 2008).

detected by infrared camera. Thus, much research effort has been concentrated in the last years on the task of automatic eye detection, localization, and tracking (Ji *et al.*, 2005; Morimoto & Mimica, 2005; Tang *et al.*, 2005; Song *et al.*, 2006).

Several authors (Pantic & Rothkrantz, 2000*a*; Sinha *et al.*, 2006; Campadelli *et al.*, 2007) have acknowledged that among other prominent facial landmarks, the eyes have several important characteristics. Thus, eyes are "stable features" in the face as they possess rich structures and exhibit high contrast. The eyes remain visible in the facial image regardless of a majority changes in the face. For example, the eyes are relatively unaffected by the presence of facial hair (*e.g.* beard, moustache, or transparent spectacles), and are little affected by small out-of-plane head rotations and degradation in the image resolution. The eyes possess unique geometric, photometric, and motion characteristics. The knowledge of the eye positions allows the rough identification of a face scale (as the distance between eyes is relatively constant from subject to subject (Farkas, 1994) and its in- and out-of-plane rotations (Hannuksela *et al.*, 2004). Further, accurate eye localization enables the identification of all the other facial features of interest. Thus the locations of other landmarks can be derived from the location of the eyes while applying, for example, a geometrical face model (Gu & Ji, 2005; Campadelli *et al.*, 2007). Finally, eye locations often serve as a criterion of successful face detection (Jesorsky *et al.*, 2002; Hamouz *et al.*, 2005; Campadelli *et al.*, 2007).

Eyebrows. Eyebrows are important facial features. In the recognition of facial identity by humans, the eyebrows make a substantial contribution to the geometric and photometric structure of the face to be recognized (Sinha *et al.*, 2006). It has been demonstrated that eyebrows are involved in a large number of facial movements (Ekman & Friesen, 1978; Ekman, 1979), therefore providing important information about emotionally and socially meaningful facial expressions.

Like the eyes, the eyebrows tend to be "stable features" in the facial image because they are relatively high-contrast and large facial landmarks. Thus the eyebrows possess the following important characteristics (Sinha *et al.*, 2006; Campadelli *et al.*, 2007). The eyebrows can survive substantial image degradations, in- and out-of-plane head rotations, and facial expressions. They can be viewed at a distance or in a low-resolution image. The eyebrows are less sensitive to shadow and illumination changes than other facial features. Further, although the eyebrows are involved in a wide range of facial movements, for example, in expression variation, the corresponding variations in the appearance of the eyebrows themselves is relatively slight compared to those observed in the eyes and mouth.

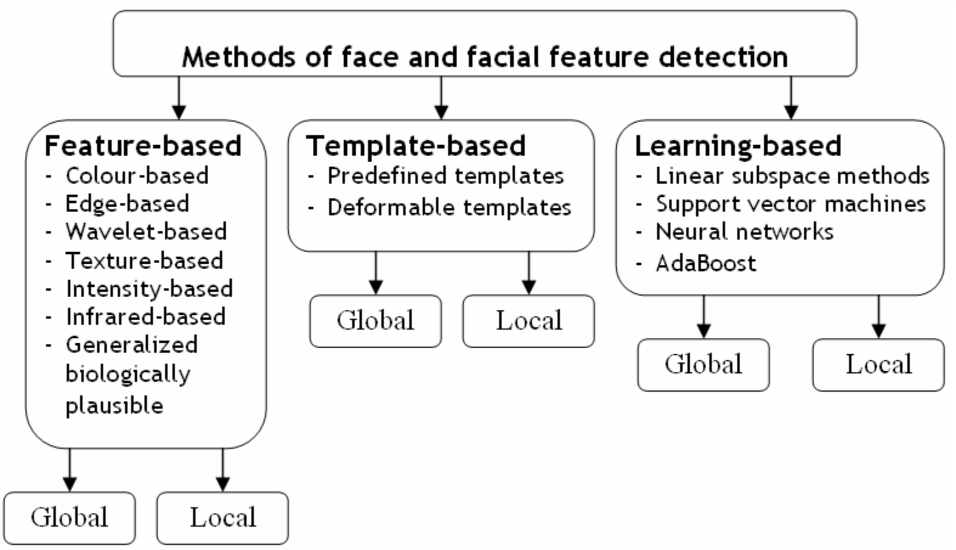
Thus the eyebrows are attractive features for detection. Face geometry is generally used to verify the positions of the eyebrows in the facial structure (Gu & Ji, 2005; Campadelli *et al.*, 2007). Some studies detect eyebrows as part of the eye regions; other studies attempt to detect eyebrows as separate facial features. However, there are some limitations in the practical use of the eyebrow landmarks. The point is that the borders of the eyebrows are not always easy to detect, especially if the eyebrows are light or the eyebrows are covered by hair (which is very often the case). Additionally, there is no common evaluation measure of eyebrow detection accuracy (Sinha *et al.*, 2006). This last issue will be discussed later in Section 2.4 of this chapter.

Mouth. The mouth is a highly deformable facial landmark prone to a wide variety of out-of-plane deformations (*e.g.* showing teeth and tongue), self-occlusions (*e.g.* bitted lips), and occlusions by structural elements (*i.e.* beard, moustache, and clothing). The mouth plays an important role in producing facial expressions and speech articulation. Automatic mouth detection can be used in facial expression analysis. Lip positions are used for lip reading to improve speech recognition, especially, in environments with high ambient noise.

Nose. The nose does not provide as high contrast and rich structure as do the eyes, eyebrows, and mouth (Wang *et al.*, 2002). It has been shown to be prone to illumination changes. The appearance of the nose changes dramatically in large out-of-plane head rotations. However, it has the important symmetrical property of having two nostrils in frontal view facial images and of being the most prominent facial feature in the profile-view images. "The nose is characterized by very simple and generic properties: the nose has a "base" the gray levels of which contrast significantly with the neighboring regions; moreover, the nose profile can be characterized as the set of points with the highest symmetry and high luminance values; therefore […] the nose tip [can be defined] as the point that lies on the nose profile, above the nose baseline, and that corresponds to the brightest gray level" (Campadelli *et al.*, 2007). These properties allow the nose to be detected from the facial image.

2.3 Approaches to Automatic Face and Facial Feature Detection

A practical solution to the problem of automatic face and facial feature detection consists of finding a representation of the face and designing an image feature detector. After the appropriate face representation has been found, the next step is to design a detector. Numerous face and facial feature detection methods have been proposed in the literature (Heisele, 2000; Pantic & Rothkrantz, 2000*a*; Hjelmas & Low, 2001; Yang *et al.*, 2002; Zhao *et al.*, 2003; Face Detection Homepage). Generally, they can be classified according to three approaches: feature-, template-, and learning-based approaches. The proposed approaches have been developed in several directions which differ from one another regarding input data requirements, computational algorithms, robustness to affecting variables, architecture, and target applications. Depending on the face representation used, the methods from each category can be further classified as global and local. Global methods are used to search for a pattern of the whole face. The methods from this category take as an input a global representation of the face. A distinctive property of these methods is that geometrical information on the facial pattern is preserved. Local methods utilize a local face representation and are applicable for the task of detecting independent facial features, for example, eyes, nose, and head outline. The location of the face can be further derived from the landmark locations. This approach has the advantage of producing the exact positions of facial landmarks which can be used in further face processing tasks. The proposed classification is illustrated in Figure 2.2. It should be mentioned, however, that a classification of face and feature detection methods into categories is a difficult task and the classification presented is largely conditional as there are also hybrid methods which are characterised by several approaches.


Methods of face and facial feature detection

Feature-based
- Colour-based
- Edge-based
- Wavelet-based
- Texture-based
- Intensity-based
- Infrared-based
- Generalized biologically plausible

Template-based
- Predefined templates
- Deformable templates

Learning-based
- Linear subspace methods
- Support vector machines
- Neural networks
- AdaBoost

Figure 2.2. Classification of face and facial feature detection methods.

Through the decades of field development, hundreds of face and feature detectors have been proposed. In this survey, the aim was to describe the current situation in the field in general rather than to provide an exhaustive description of separate methods. For this reason, only a short review for each category is given. Pros and cons are considered for each particular category. In this dissertation, the main efforts were concentrated on the development of a feature-based framework for face and facial landmark localization. For this reason, in the survey the focus is mainly on reviewing the closest references in the feature-based approach. Another specific point of interest in the survey is the consideration of the robustness property of different detection methods regarding facial changes brought about by facial muscle movements.

Feature-based Approach

The feature-based approach to object detection has been widely used in the computer vision domain (Dorko & Schmid, 2003; Heisele *et al.*, 2006; Mohan *et al.*, 2001; Schneiderman & Kanade, 2000; Ullman *et al.*, 2002; Weber *et al.*, 2000). It has been also extensively utilized for automatic face and facial feature detection. Feature-based detectors generally perform either bottom-up or top-down detection. Both detection schemes occur in two stages of analysis of the image features. Bottom-up detection starts from low-level feature analysis. The raw pixel properties of the image (*e.g.* colours or grey-scale values) are used to exploit rather abstract facial features (*e.g.* edges, lines, line intersections, points, and regions). This stage usually produces a large number of low-level features derived from the image (*e.g.* edges which correspond to a cluttered scene or elements of hair and clothing). High-level processing further combines the low-level features detected into meaningful feature formations which represent specific visual patterns for the identification of corresponding facial structures between images. The high-level stage also performs a reduction of false detections received from the low-level stage. Top-down face detection performs image processing starting from a high-level stage and is guided by the knowledge about the face. Further, a low-level feature analysis is performed and the detected face candidates are verified by a detailed analysis of local feature properties. In both approaches, in order to judge whether or not the detected image parts constitute a face, the high-level stage uses knowledge of what constitutes a human face. A constellation analysis is applied to arrange the detected feature candidates into face-like formations, for example, by statistical probabilistic models (Burl & Perona, 1996; Yow & Cipolla, 1997; Lin & Fan, 2000). Alternatively, only confidence measures received from individual low-level feature detectors are used.

Colour-based methods. The main idea that lies behind these methods is that skin colour differences between people concern intensity values rather than chrominance values. This means that a distribution of skin colour is clustered in the chrominance space. Even if skin colour distribution is considered among different ethnicity groups, it is still compact. Several colour spaces have been used for skin colour detection as, for example, (normalized) RGB, HSV (*i.e.* HSI), YCrCb, YIQ, YES, CIE XYZ, CIE LUB, CIE Lab, TSL, and UCS/Farnsworth. Various statistical analysis methods have been applied to create colour representations of the face such as histograms, Gaussian colour models, mixture colour models, and look-up colour tables. All these methods are based on collecting representative sets of skin-coloured pixels and building a skin model in a chosen colour space (Pal & Pal, 1993; Yang & Huang, 1994; Yang *et al.*, 2002; Kakumanu *et al.*, 2007).

A holistic approach to colour-based face detection is widely utilized (Chang & Robles, 2000; Martinkauppi *et al.*, 2001; Hannuksela *et al.*, 2004). It is frequently used to segment the skin-like areas of the input image or video frame, as shown in Figure 2.3. This procedure reduces the amount of information (*i.e.* pixels) to be processed at the following stages of face processing. Colour information is especially useful for face segmentation when the image is degraded due to low resolution or noise. After the skin-coloured regions have been detected, there are still false detections (*e.g.* detected hands and other skin-coloured objects) which need to be eliminated. Because colour-based methods produce many noisy candidates, further processing is required to remove small areas, fill holes in the skin-coloured regions received, fit elliptical shapes to or check on aspect ratio the detected areas. For these reasons, colour-based methods are mainly used for face and feature localization (Yang *et al.*, 2002). However they can be used for the detection task with subsequent processing of the detected regions to verify the existence of the face or facial features. The colour property of the face is also used to detect separate facial landmarks. Colour segmentation is most often applied to mouth detection (Cooray & O'Connor, 2004) as the mouth possesses a distinctive reddish colour property (Figure 2.4).

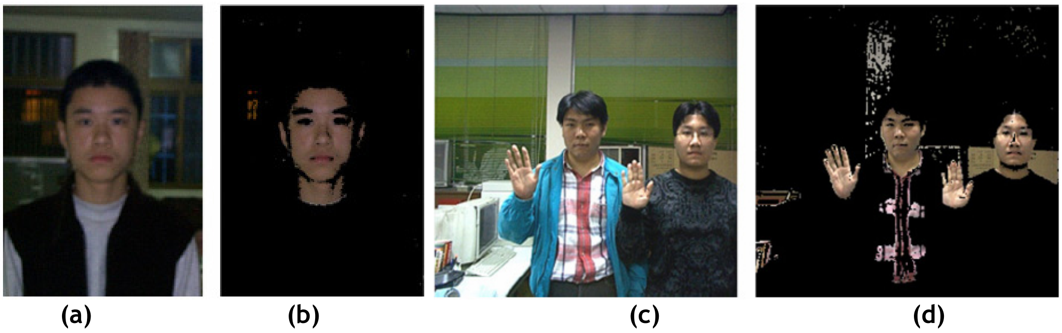


Figure 2.3. Examples of colour-based face detection: (a, c) original colour images and (b, d) the detected skin-like areas of the image. Reprinted with permission.[9]

Colour information in the image is one of the fastest to compute. For this reason colour-based detectors are applicable in real-time applications. These methods are relatively robust against head rotations and degradation in image resolution. However, they are sensitive to changes in illumination and skin tone. A selection of the appropriate skin model is one of the biggest challenges in colour-based face and feature detection

[9] Reprinted from *Pattern Recognition Letters*, *28/16*, 2007, Ch. Lin, “Face detection in complicated backgrounds and different illumination conditions by using YCbCr color space and neural network”, Fig. 5 (p. 2195), copyright © 2007 Elsevier B.V., with permission from Elsevier B.V.

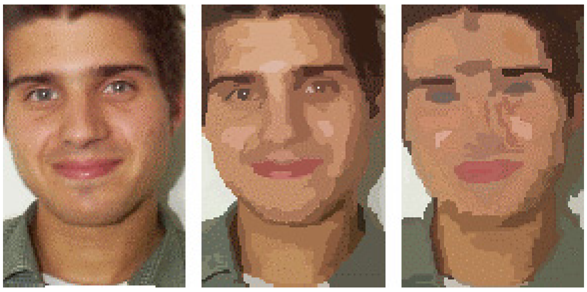


Figure 2.4. The mouth region is detected using the redness property of the segmented regions. Eyes, as areas of high intensity variance in the facial region, are searched above the detected mouth region using heuristics rules. Reprinted with permission.[10]

methods. Partial facial occlusions and facial expressions to some extent impair the performance of these methods. While a face region can be roughly estimated no matter what expression is displayed on the face, detection of separate features can be severely compromised in the presence of facial expressions. For example, bitted lips lose their reddish colour. In this case, the colour property of the lips cannot be applied for mouth detection. However, it is generally unknown what effects different facial expressions have on the performance of skin-based face and facial feature detection methods.

The advantages of colour-based methods are that they:

- are relatively easy to implement and fast to apply in real time
- allow those parts of the image to be discarded which are most likely not a face
- are usually efficient and robust with regard of head rotations, scale, moderate facial expressions, complex background, and partial occlusions

The disadvantages of colour-based methods are that they:

- require colour input data
- are sensitive to changes in illumination, skin colour, and strong facial expressions
- can produce many false positives (*i.e.* hands, neck, or skin-coloured objects in the scene) are produced in the raw output
- usually require a verification process to eliminate false positives and group colour pixels in face-like regions
- require filling holes in the skin-coloured regions received (which is a non-trivial task)

Edge-based methods. The first pioneering work on edge-based face and feature detection was accomplished by Sakai *et al.* (1971). Facial features

[10] Reprinted from *Lecture Notes in Comp. Science*, *3212*, 2004, “Facial Feature Extraction and Principal Component Analysis for Face Detection in Color Images”, Cooray S. and O'Connor N.E., Fig. 3 (p. 743), copyright © 2004 Springer-Verlag, with permission from Springer-Verlag.

were detected from the line drawings of faces from photographs. Starting from that early work, edge-based methods have been successfully applied for automatic face and feature detection. They utilize edge maps as a powerful initial representation of the image. It is assumed that the edges capture the most important aspects of the image (*i.e.* local discontinuities) filtering out less important information. Generally, edges need to be extracted, labelled, and grouped into edge patterns, which are then matched against a model in order to verify correct detections. Colour-based face detection, histogram equalization, and smoothing operators can be applied to the image prior to the edge detection step.

Many edge detection operators are available (Hypermedia Image Processing Lab gives theoretical explanations and implementation hints for many edge detectors). Among others, the most commonly cited operators are Sobel edge detector, Canny edge detector, and a variety of first and second derivatives (*i.e.* Laplacian) of Gaussian operator. Sirohey (1993) proposed a method based on the edge orientation map of the facial image. Edges were extracted and grouped into line segments. Those line segments lying on a contour of the face were combined to constitute an elliptic shape. The ellipse then was fitted on the boundary of the detected region. The best elliptic match denoted the location of the face. Chetverikov and Lerch (1993) exploited another edge orientation template matching method augmented with blob features. A face was considered to consist of two dark regions (*i.e.* eyes) and three light blobs (*i.e.* cheeks and nose tip) located in a specific spatial relationship in the image. Additionally, the orientation of edges around the blobs was checked to be consistent with the face outline and the contour of the lips. Govindaraju (1996) proposed a method based on the Marr-Hildreth edge operator (Marr & Hildreth, 1980). He used an edge map of the original facial image to form a feature vector from the extracted specific edge groups (*i.e.* lines, arcs, and their combinations). The oriented elements detected were labelled as belonging to the "left side", "hairline" or "right side" of the face. These edge groups were matched against the edge orientation template of the face that was built on the basis of the gold standard ratio[11]. A cost function estimated likelihoods for each face candidate. An interesting approach was proposed by Yow and Cipolla (1997) in which they considered a face as a pattern of two “edge-free” and six “edge-full” features. “Edge-free” features corresponded to cheeks and “edge-full” features corresponded to eyebrows, eyes, nose, and mouth all of which consisted of primarily horizontally oriented edges. Recently, Fröba and Küblbeck (2000; 2002) proposed another matching method to detect a face from the image. A facial template was constructed on the basis of local

[11] The height of the face divided by its width equals φ, where $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.6180339887$.

oriented edges from the areas of the eyes, nose, and mouth (Figure 2.5). Using this template, a face was searched for in the image. Similar approaches have been used to detect separate facial landmarks (Low & Ibrahim, 1997; Choi *et al.*, 1999).

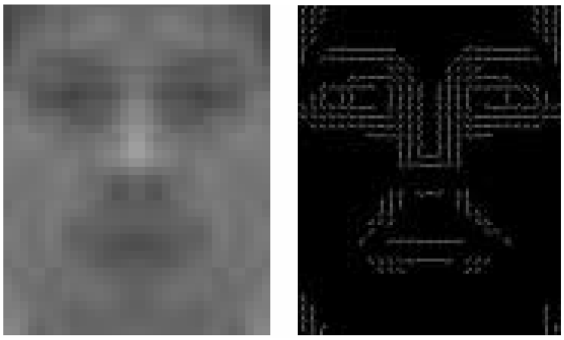


Figure 2.5. The left image shows a smoothed original facial image. The right image demonstrates orientation template of the face constructed on the basis of the edge map of the image. Reprinted with permission.[12]

Kawaguchi and Rizon (2003) located eye pupils using intensity and edge information. For this purpose they included in their algorithm Sobel edge operator and feature template matching. In combination with other information (*e.g.* colour and intensity), edges have been widely utilized for purposes of mouth and eye corner detection. Several algorithms based on a search for local oriented edges constituting lines and line intersections were developed specifically for this purpose. Herpers *et al.*, (1996) utilized the first and second derivatives of Gaussian operator to detect oriented edge and line segments from a facial image. Using models of the eyes and mouth they performed template matching to find oriented elements constituting eye and mouth corners. A context-based feature search followed to locate eyebrows and nose. Pahor & Carrato (1999) proposed a method to detect the mouth corners of a speaker from videophone sequences. Campadelli *et al.* (2007) applied edge detection operators to detect rough eye and mouth locations as a preprocessing step for snake initialization. These methods were proved to be robust with regard to facial expressions (*e.g.* mouth was detected as closed or opened with teeth visible (Pahor & Carrato, 1999; Pantic & Rothkrantz, 2000*b*; Campadelli *et al.* 2007).

Because edges represent local discontinuities of the visual scene, edge-based methods have been seen to be largely invariant to skin colour and shallow shading gradients resulting from small variations in illumination and out-of-plane head rotations. The main difficulty in applying edge-

[12] Reprinted from *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'02)*, 2002, "Robust Face Detection at Video Frame Rate Based on Edge Orientation Features", Fröba B. and Küblbeck C. Fig. 3 (p. 3), copyright © 2002 IEEE, with permission from IEEE.

based methods for the task of face and feature detection is that it is not trivial to compose edges and make sense of the edge patterns detected. The performance of the edge-based detectors depends heavily on the threshold values used for edge grouping. In a cluttered scene it is most likely to yield a lot of false detections. It is difficult to define which edges belong to the face. For this reason, edge-based methods are usually applied under constrained conditions or after the image has been preprocessed (*e.g.* by a colour-based face detector). Alternatively, additional information is required, for example, if it is known in advance that the face to be detected has a frontal view. Edge-based methods are mainly used for face and feature localization (Yang *et al.*, 2002). Generally, local methods demonstrate better performance under varying conditions than global ones.

Advantages of edge-based methods are that they:

- work with grey-scale images
- are relatively simple and easy to implement
- demonstrate good empirical results under constrained conditions
- are useful for eye and mouth corner detection from images with facial expressions if the rough position of the feature is known
- perform data compression at the very early stage of image processing

Disadvantages of edge-based methods are that they:

- are sensitive to complex background
- are sensitive to occlusions, facial expressions, changes in lighting, and out-of-plane head rotations if applied globally
- require a selection of threshold values used for edge detection and grouping
- may be computationally expensive and slow if applied at multiple orientations and resolution levels

Wavelet-based methods derive multi-resolution and multi-orientation wavelets from the image. The wavelets possess the important property of describing faces in terms of luminance changes at different frequencies, positions, and orientations. This property allows the characteristic patterns for face and feature detection to be composed. On the other hand, the use of wavelets for the task of face detection is motivated by strong biological analogies. Wavelet representation constitutes a good model of the responses of the simple cells in the visual cortex (Orban, 1984; Hubel, 1995; Ullman *et al.*, 2002). Figure 2.6 presents an example of Gabor wavelet-based face representation. One of the famous wavelet-based face and feature detection methods was that proposed by Wiskott *et al.*, (1997). This method, called elastic bunch graph matching, located facial features using object adopted graphs. The face geometry was encoded by a structure of the graph. The information on local feature points in each node of the elastic graph was represented by Gabor wavelets. A feature vector was generated by concatenating all the wavelet coefficients together. The facial features were extracted by maximizing a similarity between the novel image and the model graphs.

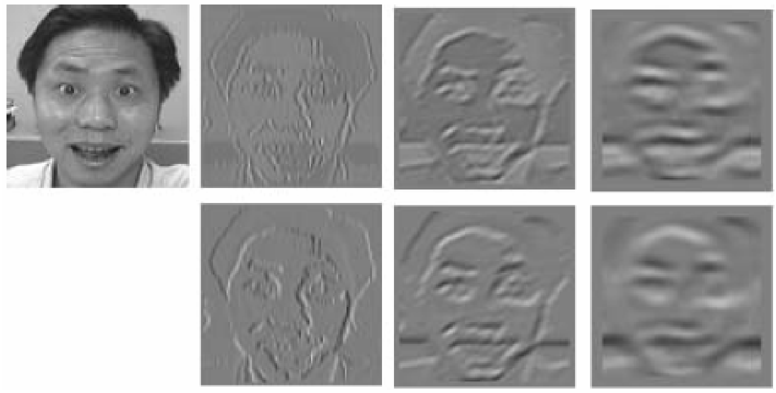


Figure 2.6. A Gaborface representation of the facial image demonstrates the convolution results of the image with three Gabor kernels. Reprinted with permission.[13]

Wavelet-based methods can be also applied locally in the detection of separate facial landmarks. In this case, a multi-resolution and multi-orientation landmark representation is formed from the wavelet coefficients derived from the regions of the landmark as shown in Figure 2.7 (Huang & Wechsler, 1999, Feris *et al.*, 2002; Fasel *et al.*, 2005; Song *et al.*, 2006; Campadelli *et al.*, 2007).

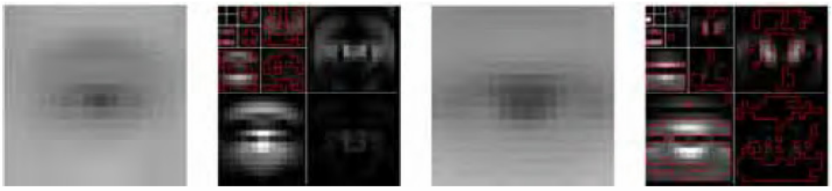


Figure 2.7. From left to right: “mean eye” pattern; its wavelet decomposition; “mean eye” pattern at higher resolution; and its wavelet decomposition. The selected features of the two eye patterns are shown as red contours. High intensities correspond to strong edges and low intensities indicate uniform regions (Campadelli *et al.*, 2007). Courtesy of P. Campadelli, R. Lanzarotti, and G. Lipori. Reprinted with permission from P. Campadelli.

In general, Gabor-based detection methods have been demonstrated to be effective and robust under many unconstrained conditions. The reason for this is that Gabor-based face representation eliminates variability in the image due to variations in lighting conditions, contrast, and small head rotations. This representation is robust against small shifts and deformations in the facial image due to facial expressions (Tian *et al.*, 2002;

[13] Reprinted from *Machine Vision and Applications*, *16*, 2005, “Information Extraction from Image Sequences of Real-World Facial Expressions”, Gu H. and Ji G., Fig. 4 (p. 107), copyright © 2004 Springer-Verlag, with permission from Springer-Verlag.

Gu and Ji, 2004). The complexity of a scene impairs the performance of these methods, thus preprocessing is usually performed prior to the detection step. Although the wavelet representation of a face has been widely adopted, it is computationally expensive to convolve the input image or video frame with multi-banks of wavelet filters. Wavelet-based methods are mainly used for face and feature localization (Yang *et al.*, 2002).

Advantages of wavelet-based methods are that they:

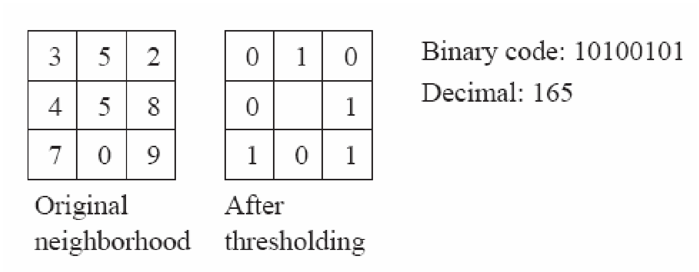
- are relatively effective and efficient under changes in lighting, small head rotations, and facial expressions
- simulate low-level mechanisms of biological vision (*i.e.* local orientation sensitivity of neurons of the primary visual cortex)

Disadvantages of wavelet-based methods are that they:

- are usually computationally slow to be applied in real-time
- are sensitive to cluttered scenes, occlusions, large out-of-plane head rotations, large changes in lighting, and strong facial expressions

Texture-based methods. The texture property of the face has also been used for automatic face and feature detection. The assumption is that human faces have a distinctive texture that can help to distinguish faces from objects of other classes. These methods began with work by Augusteijn and Skujca (1993). The general approach proposed is that texture is first modelled by statistical methods. Next, a classifier is trained to discriminate between face and non-face texture patterns. Augusteijn and Skujca used a spatial grey-level difference-based method that utilized colour information in combination with second-order statistical features which model texture patterns. The texture was classified into three classes (*i.e.* face, hair, or others) by a neural network. In other works, high-order autocorrelation coefficients were used to construct texture models (Hotta *et al.*, 1998; Kurita *et al.*, 1998; Popovici & Thiran, 2001).

Another texture-based face detection method utilized local binary pattern (LBP) operator (Ojala *et al.*, 2002; Hadid *et al.*, 2004; Ahonen *et al.*, 2006). In this method, a facial image was first divided into a set of blocks. Next, from each block the LBP feature histograms representing texture contents within these regions were computed. In the next step, all intermediate LBP feature histograms were concatenated into a single histogram. Figure 2.8 illustrates the process of calculating the local LBP code. Pixels from each local region (coded from 0 to 256) were labelled according to the value of the central pixel. The result of the labelling was considered a binary number which coded a local LBP texton at a given location in the image. The proposed method achieved good results on face datasets in unconstrained conditions (*i.e.* changes in illumination, head rotations, occlusions, *etc.*) and facial expressions (Zhao & Pietikäinen, 2007; Feng *et al.*, 2005).



3	5	2
4	5	8
7	0	9

Original neighborhood

0	1	0
0		1
1	0	1

After thresholding

Binary code: 10100101
Decimal: 165

Figure 2.8. The process of pixel labelling in the image neighbourhood of 3 by 3 pixels to calculate a local LBP code at this pixel position. Reprinted with permission.[14]

Advantages of texture-based methods are that they:

- work with grey-scale images
- usually require high resolution images with visible texture patterns
- demonstrated good empirical results under constrained conditions
- are relatively simple and easy to implement in real time

Disadvantages of texture-based methods are that they:

- are sensitive to large changes in illumination, large out-of-plane head rotations, occlusions, strong facial expressions, and complex background

Intensity-based methods. Intensity properties of the image have also been utilized to detect faces and features globally or locally. Yang and Huang (1994) introduced a rule-based multi-resolution method to locate a face from the image based on the observation that humans can detect faces at a very low resolution. They used the assumption that facial landmarks differ from the rest of the face because of their low brightness. Low-resolution images were used to detect all possible face candidates as bright areas of a predefined size. Increasing the image resolution, they detected eyes, nose, and mouth as dark regions inside the face candidates. High-resolution images were used to analyse the received feature candidates in detail. A set of rules was applied to the facial structure of images at each resolution level to reduce the amount of information to be processed at a higher level of resolution. This method was elaborated by Lv (2000) and Kotropoulos and Pitas (1997). Sobottka and Pitas (1997) analysed a topographic grey level relief of the face region. After the darkest areas in the face were detected, a geometry face model was used to capture the best face-like constellation of the landmark candidates detected. Mäkinen and Raisamo (2002) applied a similar method in which a topographic grey level relief of a face region was used to construct intensity profiles of the facial pattern (Figure 2.9). The skin-coloured face candidates detected were verified by ellipse fitting procedure and detection of facial landmarks as the darkest areas inside of the face candidate.

[14] Reprinted from *Real-Time Imaging*, 9/5, 2003, Mäenpää T., Turtinen M., and Pietikäinen M., "Real-Time Surface Inspection by Texture", Fig. 1 (p. 290), copyright © 2003 Elsevier, with permission from Elsevier.

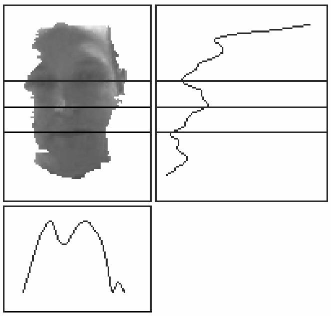


Figure 2.9. The left image shows the colour-segmented facial region. The right graph shows the y-axis grey level relief of the whole facial region. The bottom graph shows the x-axis grey level relief of the eye regions. Reprinted with permission.

Other methods utilize the distinctive intensity properties of the face to detect facial features. For example, the intensity map of the image is used to detect bright landmarks like the nose as the lightest area in the image (Hoogenboom & Lew, 1996; Campadelli *et al.*, 2007). Figure 2.10 shows the detection results for the tip of the nose.

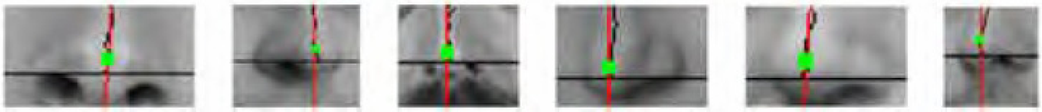


Figure 2.10. Examples of nose processing (Campadelli *et al.*, 2007). The horizontal line indicates the base of the nose, the dots along the nose are the points of maximal symmetry along each row, the vertical line approximates those points, and the rectangular marker indicates the tip of the nose. Courtesy of P. Campadelli, R. Lanzarotti, and G. Lipori. Reprinted with permission from P. Campadelli.

Intensity-based methods are useful for purposes of rough estimation of face and feature locations, for example, in those cases when the image has a low resolution, is corrupted by camera noise, or a face is shown at a distance (Campadelli *et al.*, 2007). Intensity-based methods are robust to facial expressions (*i.e.* the face appears as bright area and facial features appear as dark areas in the facial region even if a strong facial expression is shown). Another advantage of these methods is that they do not require high-resolution images and some blurring effect of the image is even desired. However, intensity information has been shown to be prone to large changes in lighting, large head rotations, and complex background. The performance of intensity-based detectors depends largely on the threshold values selected for binarization of the intensity maps. These methods are mainly used for face and feature localization (Yang *et al.*, 2002).

Advantages of intensity-based methods are that they:

- have low computational requirements and can be applied in real time
- work with grey-scale low-resolution images

- are robust regardless of small changes in lighting, small head rotations, moderate facial expressions, scale, and camera noise

Disadvantages of intensity-based methods are that they:

- usually provide rough detection results which need to be refined by other methods
- require a selection of threshold values used for binarization of the intensity map
- are sensitive to large changes in lighting, large out-of-plane head rotations, occlusions, strong facial expressions, and complex background

Infrared-based methods use a phenomenon of reflection of the infrared light by the eye pupils. These methods are used to detect pupil locations first. Following this, other features can be extracted next. Usually, some thresholding process precedes the detection step to make the pupils the darkest or lightest regions in the image. Gu and Ji (2005) applied light- and dark-pupil techniques to detect eyes (Figure 2.11). They used Gabor eye region decomposition and head movement prediction on the basis of the detected eye pupils to extract a number of characteristic points from the face. Infrared-based methods provide a powerful means to accurately detect eye pupils regardless of wide variability in the affecting variables.

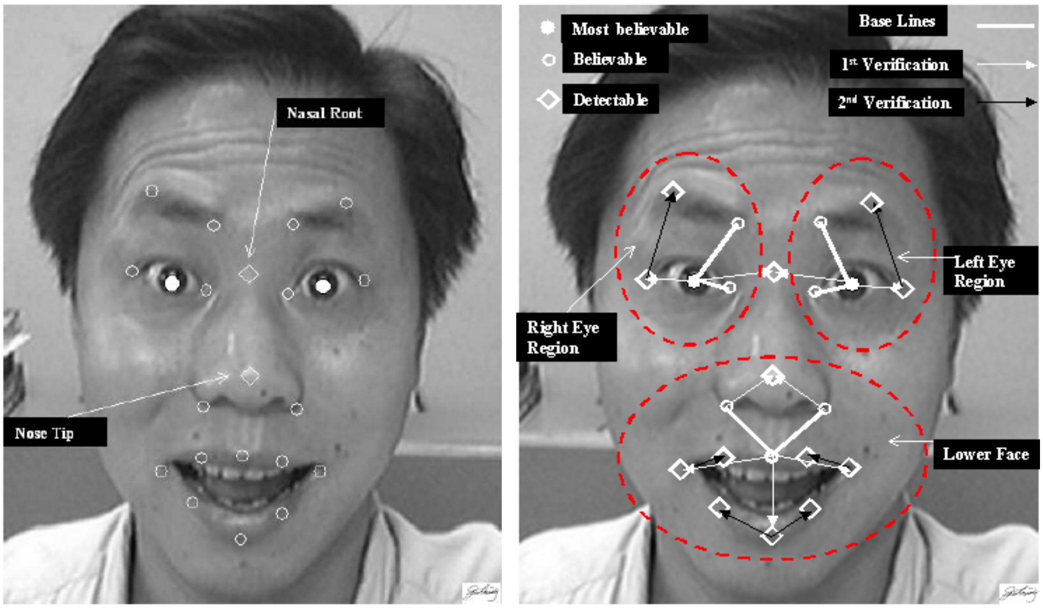


Figure 2.11. An example of infrared-based eye detection. The left image shows the detected eye pupils (filled circles) and feature points (unfilled circles). The right image shows the verified feature locations. Reprinted with permission.[15]

[15] Reprinted from *Machine Vision and Applications*, *16*, 2005, "Information Extraction from Image Sequences of Real-World Facial Expressions", Gu H. and Ji G., Fig. 3 (p. 107) and Fig. 5 (p. 109), copyright © 2004 Springer-Verlag, with permission from Springer-Verlag.

The face can be effectively detected by these methods on the basis of eye locations if the eye pupils are visible. Scale, head rotations, and complex background do not affect the performance of these methods. However, the detection may fail if the eyes are occluded by structural components or eyelids during facial reactions.

Advantages of infrared-based methods are that they:

- work with colour and grey-scale images
- are relatively simple and easy to implement in real time
- are effective and robust regardless of variability of conditions

Disadvantages of infrared-based methods are that they:

- require the use of infrared imaging technique
- require a selection of parameters for binarization of infra-red camera image
- require the eyes to be open, which is not always the case in variations in expression

Generalized biologically plausible methods. As described, visual features like colours, textures, edges, symmetry properties, and intensity values are derived and processed at the early stages of the biological visual system. These low-level pre-attentive mechanisms help the visual system to concentrate on the important parts of the image while discarding the less important ones. A number of computer vision researchers (Rybak *et al.*, 1990; Reisfeld, 1993) have suggested that automatic face and feature detection should start from an analysis of the pre-attentive low-level generalized properties of the image. In their later work, Reisfeld and Yeshurun (1998) further developed this idea to utilize a symmetrical property of facial features. Their method was based on the analysis of edge and intensity maps of the image guided by knowledge of the face structure and symmetry of the facial landmarks.

Rybak *et al*. (1990) utilized the analysis of edge structures that simulated a property of orientation selectivity of the neurons in the primary visual cortex. This work was extended to construct the attentive behavioural model of vision (BMV) (Shaposhnikov *et al*., 2002; Rybak *et al*., 2005; Shevtsova *et al*., 2007). In the model, the mechanisms of fixation point selection strategy of the human eyes were exploited to detect prominent facial features. The simulated mechanisms of visual attention guided a selection of low-level features such as edges at particular orientation and intensity values. The difference between two multi-orientation Gaussians was used to extract local oriented edges from the image at several levels of resolution. The orientation of the edge at a given position of the image was considered in a context of orientations of edges in the neighbouring image parts. The BMV model performed "a switch of attention" to the neighbouring part of the image in which the orientation of the edge was maximally different from the preceding one. Herpers (1995) proposed a similar method based on a simulation of the processes of active vision. A multi-orientation and multi-resolution Gaussian edge detector was used to extract local oriented edges. Using a symmetry property of the eye pair, the eye regions were detected at the low resolution. The positions of the eyes further guided the search for nose and mouth features. Recently, an extension of this work with Gabor wavelet-based face representation at multiple scales and orientations was reported by Smeraldi *et al.* (2000).

Generalized feature-based face and feature detection methods accumulate knowledge on feature-based image processing and take advantage of many image features simultaneously. They use knowledge about how the visual system processes this information. For this reason they have been shown to be robust with respect to many variations in facial appearance caused by changes in lighting, facial expressions, and in-plane head rotations.

Advantages of generalized methods are that they:

- accumulate the advantages of feature-based face and feature detection methods
- have the important property of being biologically plausible and simulate the low-level pre-attentive mechanisms of biological vision

Disadvantages of generalized methods are that they:

- have the general disadvantages of feature-based detection methods

Template-based Approach

The methods in this category apply a correlation between facial template(s) and the input image or video frame. The procedure consists of two steps. In the first step, a facial template has to be created (usually manually). In the next step, a search is performed for the area in the image or video frame that resembles the template(s). After that, the existence of the face or facial feature in the image is declared in the area that gained maximum correlation. The structure of the facial templates can be predefined (*i.e.* permanent) or deformable. A description for each template category is given below.

Predefined templates. Predefined templates were generally among the first methods of automatic face and feature detection. A classical example of a predefined template is the ratio template proposed by Sinha (1994). The template consisted of 16 adjacent regions and 23 relations between these regions. The relations of the template denoted the direction of changes in the brightness of the neighbouring facial regions. In this work, a matching strategy was applied to the input image in order to derive the best face candidates. The ratio template used in this work is shown in Figure 2.12.

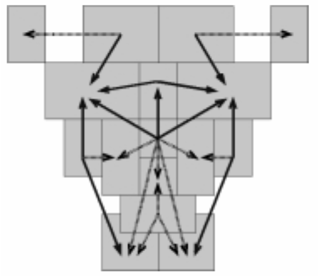


Figure 2.12. Facial ratio template based on relative brightness of 16 neighbouring regions (grey boxes) and 23 relations between regions (arrows) (Sinha, 1994). Reprinted with permission.[16]

The predefined templates are usually easy to construct and compute. It makes sense to use the predefined templates for face and feature detection under strictly controlled conditions (*i.e.* constant lighting, restricted head movement, simple background, neutral (or known) expression, *etc.*). However, if these conditions fail, the detection rates decrease significantly. The deteriorating effect of facial expressions on face detection made by matching a global predefined template against the image can be illustrated as follows. Let the template of the face consist of dark areas for eyebrows, eyes, nose, and mouth as shown in Figure 2.13. It is seen that the expression of surprise leads to inconsistency between the template and the locations of the eyebrows and mouth. The template can be constructed, for example, from the edge or intensity information. Still, the presence of a strong facial expression, for example, surprise, leads to noticeable inconsistency between the positions of landmarks in the template and the image. The use of predefined templates locally for the detection of separate facial landmarks (Evreinova *et al.*, 2006) partly solves the problem, as separate facial parts exhibit less variability than the pattern of the whole face.

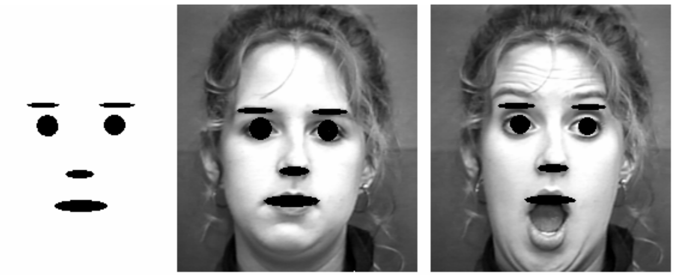


Figure 2.13. The global facial template consists of dark areas which correspond to the prominent facial landmarks. It can be seen that the expression of surprise leads to inconsistency between the template and the locations of the eyebrows and mouth. The image is from the Cohn-Kanade AU-Coded Facial Expression Database (Kanade et al., 2000). Reprinted with permission.

Advantages of predefined templates are that they:

- work with grey-scale images
- are relatively simple and easy to implement

[16] Reprinted from *Proc. National Conf. Artificial Intelligence*, 1998, "Eye Finding via Face Detection for a Foevated, Active Vision System", Scassellati B., Fig. 3 (p. 971), copyright © 1998 Association for the Advancement of Artificial Intelligence (AAAI), with permission from AAAI.

- are effective and efficient under very constrained conditions

Disadvantages of predefined template are that they:

- posses a difficult problem of enumeration of templates for different conditions (*i.e.* in- and out-of-plane head rotations, facial expressions, and scale) if applied globally
- can be computationally expensive and slow when applied in real-time (*i.e.* it is time-consuming to perform cross-correlation on multiple scales)

Deformable templates. Because faces constitute a class of highly deformable objects, it is more beneficial to use deformable facial templates which can adapt to facial distortions (Yuille *et al.*, 1992). Recently, extensive work has focused on a deformable property of the face. Active contour models or snakes were first proposed by Kass *et al.* (1988) and further developed in many studies (Hamarnen, 2000; Perlibakas, 2003; Campadelli *et al.*, 2007). Figure 2.14 illustrates examples of the results of the snake-based face and feature detection.

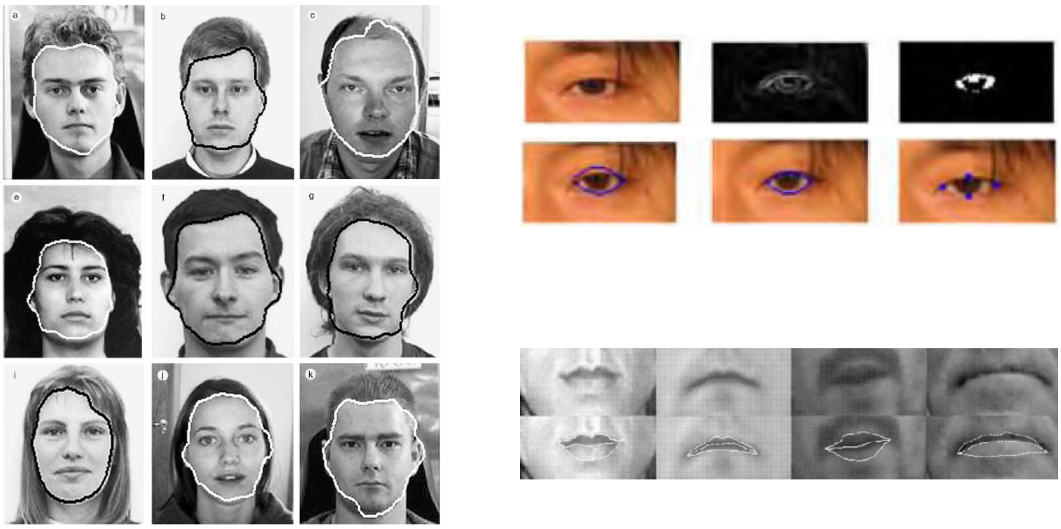


Figure 2.14. The left image shows the results of snake-based face contour detection. Reprinted with permission.[17] The top-right image shows snake-based eye detection (Campadelli *et al.*, 2007). From left to right and top to down: eye subimage; edge map; binarized map; initial template position; final template position; and detected eye points. Courtesy of P. Campadelli, R. Lanzarotti, and G. Lipori. Reprinted with permission from P. Campadelli. The bottom-right image shows snake-based mouth detection. Images were collected from http://www.cvc.uab.es/~jordi/cares.html on July 14, 2008. Courtesy of P. Radeva, A. Martínez, and J. Vitrià. Reprinted with permission from P. Radeva.

Active appearance models (AAMs) are another example of deformable templates. AAMs were first proposed by Lanitis *et al.* (1995) and further developed by Cootes *et al.* (2001). In AAMs, the characteristic points of the

[17] Reprinted from *Pattern Recognition Letters, 24/16*, 2003, Perlibakas V., "Automatical Detection of Face Features and Exact Face Contour", Fig. 2 (p. 2983), copyright © 2003 Elsevier, with permission from Elsevier.

face are annotated manually for each face in the training image set. The shape and grey-scale variation properties of the faces from the training set are learned through statistical analysis. The AAM mesh is built next on the basis of the learned shape and texture variations of the face. In the testing phase, the matching is performed against the image or video frame. It involves a search for model parameters which minimise the difference between the image and a synthesised model example projected onto the image. The potentially large number of parameters makes this a difficult problem. Figure 2.15 illustrates the steps described in the AAM-based detection process.

There are other examples of deformable templates such as active shape models (Cootes *et al.*, 1995), direct appearance models (Hou *et al.*, 2001), morphable models (Jones & Poggio, 1998), constrained local models (Cristinacce & Cootes, 2006), and active blobs (Sclaroff & Isidoro, 2003).

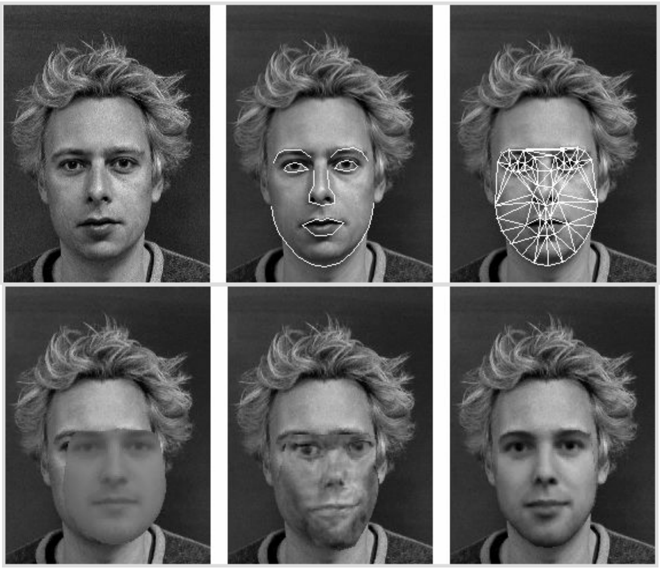


Figure 2.15. Face and feature detection by active appearance models (AAM). From left to right and top to bottom: original facial image; segmented face and feature contours; resulting AAM mesh; initial AAM model; AAM model after 2 iterations; and converged AAM model after 12 iterations. The images were collected from http://www2.imm.dtu.dk/~aam on July 14, 2008. Reprinted with permission from M. Stegmann.

Deformable templates such as snakes have been demonstrated to be robust regardless of facial deformations (Hamarnen, 2000; Campadelli *et al.*, 2007). Thus, face and features can be successfully detected in spite of wide variations in expression when the eyes are closed or semi-closed and the mouth is open and the teeth are visible. Other deformable templates might suffer from strong facial expressions as AAA-based methods do. A limitation of snake-based detectors is that they have to be initialized near the face or feature in the image. Otherwise they may fail to get a right start. Another limitation of deformable templates is that they usually require high-resolution images and are not easily achievable in real time. However, once the face is detected, it is then possible to track it on the basis of the features detected (Toennies *et al.*, 2002). Marked changes in lighting, out-of-plane head rotations, and occlusions greatly impair the performance of these methods. Therefore a preliminary face or feature detection is required to perform the detection in cluttered scenes. Template-based methods are mainly used for feature localization (Yang *et al.*, 2002).

Advantages of deformable templates are that they:

- work with grey-scale images
- are effective and efficient in constrained conditions
- can precisely detect facial features with expression variations

Disadvantages of deformable templates are that they:

- need to be initialized near the face or feature in the image
- are sensitive to large changes in lighting, large head rotations, occlusions, complex background, and strong facial expressions
- require high-resolution images

Learning-based Approach

Another class of face and feature detectors is represented by learning classifiers. These methods learn the appearance of the face and facial parts from examples. Face or feature models are constructed from the training image set and further searched for in the image. This way, no knowledge of facial structure is required. Usually, the bigger the training set, the better differentiation between face and non-face classes can be achieved. The training set should cover all variability of facial appearances. A windowing technique is usually applied in which a classifier scans the image or video frame in a window of a predefined size. The window size of the classifier is smaller than the size of the image, and a classifier therefore produces multiple output values at each scanning position. When the output of the classifier is greater than some threshold, the part of the image is labelled as being a member of the face class. The detection at multiple scales and orientations is performed by scaling and rotation of the input image. Because the whole image is fed to the classifier, the learning classifiers need to deal with the problem of dimensionality reduction. They usually apply a preprocessing step to reduce a dimension of the input data and select optimal modal parameters in the received low-dimensional space.

Linear Subspace Methods assume that any face can be represented as a linear combination of all other faces from the face space. Examples of methods from this category are principal component analysis (PCA), Gaussian component analysis, independent component analysis, factor analysis, projection pursuit, and symplectic maps to name few. For a recent review on these methods see (Deco & Obradovic, 1996).

PCA (Turk & Pentland, 1991; Belhumeur *et al.*, 1997; Gu *et al.*, 2001) is perhaps one of the most practical and systematic methods exploiting the idea of linear subspaces. PCA-based methods project the face space into a lower-dimensionality feature space by PCA. The received set of standardized face components (*i.e.* eigenvalues and eigenvectors) derived by PCA from a large training set of facial images is called eigenfaces. Eigenfaces provide a compact representation of the face to be used for face detection. The process of deriving eigenfaces proceeds as follows. First, a training dataset of facial images should be constructed. The training dataset is usually taken in the same lighting conditions, head poses, simple background, and moderate expressions and normalized by scale (*i.e.* distances between eyes), face location, and head rotation. After the training set is complete, the mean is subtracted and the covariance matrix is calculated from it. Eigenvalues and eigenvectors are further derived from the covariance matrix of the probability distribution of the high-dimensional vector space of the training dataset. Because the dimensionality of the received eigenvectors is high, only few eigenvectors with large eigenvalues are typically selected. The received eigenfaces appear as light and dark areas arranged in a specific pattern (Figure 2.16).

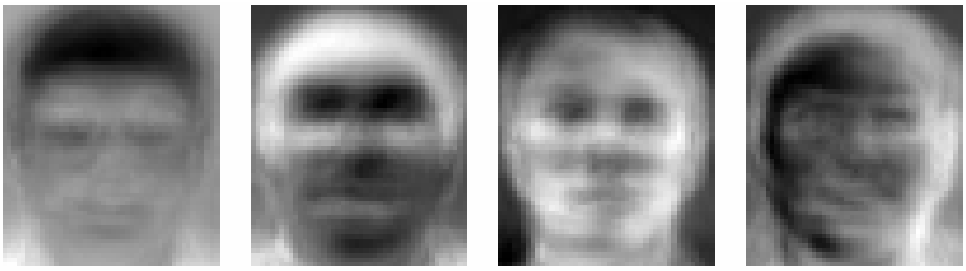


Figure 2.16. Examples of eigenfaces. © 2002 AT&T Laboratories, Cambridge. Reprinted with permission from AT&T Laboratories.

This pattern is used to differentiate between face and non-face object classes by calculating the distance from the face space. If a distance from the face space is less than a threshold, the existence of a face in the image is declared. The same idea was extended to detect separate facial landmarks by Moghaddam and Pentland (1994). They obtained eigenvalues and eigenvectors from a representative training set of each feature (*e.g.* eigeneyes, eigennose, and eigenmouth).

A difficulty of using PCA as a face detector is due to the fact that the scale and position of the face in the image is usually unknown. As a result of this, the search space is rather large and false detections are likely. Cooray and O'Connor (2004) resolved this problem by performing PCA on a region derived from the automatically detected locations of eyes and mouth. The distance between the eyes denoted a rough scale and position of the face for a normalized search space. Another drawback of this method is the requirement of face normalization by scale, location, and head rotation in the training set. Pronounced facial expressions usually lead to a misalignment error because the correspondence between landmark positions in the images of different expressions is poor. One possible solution would be to construct models (*i.e.* eigenspaces) of faces for each particular facial expression (Frank & Nöth, 2003). However, the presence of between- and within-individual differences of facial expressions would still cause difficulties in the construction of normalized spaces for each type of expressive face. PCA-based methods are used for face and feature localization and detection (Yang *et al.*, 2002).

Advantages of linear subspace methods are that they:

- work with grey-scale images
- achieve good results for faces in constrained conditions

Disadvantages of linear subspace methods are that they:

- may be computationally expensive and slow
- require a large amount of training data
- require normalization of testing data by scale, position, orientation, and expression
- have a dimensionality reduction problem in the extraction of eigenvalues from the training set
- require a search over space and scale if the face scale is not known in advance
- are sensitive to alignment error, background complexity, scale, head pose, change in illumination, and facial expressions

Neural Networks (NNs) approach the face and feature detection problem as a pure pattern recognition task. The literature includes numerous NNs which differ from one another in terms of architecture, number of hidden layers and receptive fields, learning algorithms (*e.g.* AND, OR, voting, or separate arbitration networks), and others. One of the most significant works on NN was perhaps that by Rowley *et al.* (1998*a*). They used a retinal-based multi-layer perceptron to learn a discriminant function from the training set. They used three types of receptive fields. The first receptive field resulted from a division of the facial image of size 20 by 20 pixels into four 10x10 subregions. The remaining two receptive fields were obtained from 16 subregions of size 5x5 pixels and six overlapping subregions of size 20 by 5 pixels. Each of these subregions corresponded to one hidden unit. They applied different learning algorithms to achieve the best performance. The extension of this work was presented in (Rowley *et al.*, 1998*b*) in which the architecture of the system was improved by an NN router. The NN router detected the head rotation in the input image and passed on the image to one of two NNs trained either in frontal or profile view facial images. Féraud *et al.* (2001) proposed an NN based on a constrained generative model network. The idea was to train the NN to perform a nonlinear dimensionality reduction (*i.e.* a nonlinear PCA) and then to measure the reconstruction error of each input image. A low error would signal the presence of a face. Roth *et al*. (2000) proposed a method called sparse networks of winnows (SNoWs). SNoW was a sparse network of linear functions which were defined over a pre-defined or incrementally learned feature space. This system was specifically designed for the learning of a very large amount of features. To detect a face, SNoW used two NN consisting of 2 linear thresholding units which operated on Boolean features. The Boolean features encoded positions and intensity values of pixels in the image. To perform analysis at multiple scales, the mean and variance values of the pixels at each scale were additionally calculated. The two linear units were separated from each other and sparsely connected over the Boolean feature space. The training procedure promoted or demoted the weights corresponding to the features in accordance with the classification success of the current training example.

NNs usually achieve a high performance if the training set is representative enough to cover a large range of facial variations (Garcia & Delakis, 2002). For example, a SNoW system was tested on a very large representative set of facial images with different lighting, facial expressions, and head rotations. The performance was sufficiently high for all these conditions. In contrast to PCA-based methods, NNs achieve greater invariance regardless of scale, orientation, and resolution. NN-based methods are mainly used for face and feature detection (Yang *et al*., 2002).

Advantages of NN-based methods are that they:

- work with grey-scale images
- use powerful machine learning algorithms
- are effective and efficient in many conditions including changes in lighting, head rotations, occlusions by eye-glasses, and facial expressions, if the training set includes all these variations
- are fast to apply in real time

Disadvantages of NN-based methods are that they:

- have a problem with the selection of training data and training procedure
- need many positive and negative examples
- are time consuming in training and testing
- require a search over space and scale if the face scale is not known in advance
- require the right parameters (*e.g.* number of layers, hidden units, and learning rate) to be selected

Support Vector Machines (SVMs) are hyperplane classifiers based on the concept of decision planes that define decision boundaries. From the training set of face and non-face images, an SVM classifier learns how to construct a separating hyperplane that maximizes the distance between these two data sets. In the training phase, the SVM classifier constructs a separating hyperplane in an n-dimensional space of the input image. In order to achieve good separation, the hyperplane is constructed so that it has equidistances to the neighbouring data points of both classes. These closest points are called support vectors. The greater the distance, the better the generalization error of the classifier will be. To perform a nonlinear separation, the input space is mapped onto a higher dimensional space using Kernel functions.

Osuna *et al.* (1997) first proposed applying SVM to the face detection task. This approach has subsequently been further developed (Heisele *at al.*, 2003) and applied to the detection of facial features (Heisele *at al.* 2001*a*; 2001*b*; 2006; Bileschi & Heisele, 2003). Campadelli *et al.* (2007) extended this method to detect eye regions in a coarse-to-fine manner. They used two SVM eye detectors applied to eye representations consisting of optimally selected wavelet coefficients. On the coarse level, rough eye locations were detected by an SVM trained on a representative set of eye images (Figure 2.17, top row). On the next level, the detection output from the coarse eye detector served as an input for the high-level eye detector that improved the localization precision (Figure 2.17, bottom row). Other facial features were defined on the basis of the eye locations detected. SVM-based methods are mainly used for face and feature detection (Yang *et al.*, 2002).

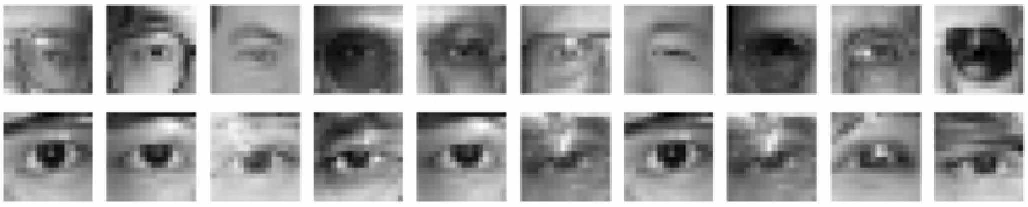


Figure 2.17. Coarse-to-fine eye detection (Campadelli *et al.*, 2007). The top row shows low-level eye region detection. The bottom row shows high-level eye position refinement. Courtesy of P. Campadelli, R. Lanzarotti, and G. Lipori. Reprinted with permission from P. Campadelli.

Advantages of SVM-based methods are that they:

- work with grey-scale images
- use powerful machine learning algorithms
- if trained on a representative image set, are effective and efficient under many conditions including changes in lighting, occlusions, head rotations, and facial expressions
- are usually fast for application in real time

Disadvantages of SVM-based methods are that they:

- need many positive and negative examples
- have a problem with the selection of training data and training procedure

- are time consuming in training and testing
- require the selection of the right parameters (*e.g.* kernels have to be known in advance)
- require searching over space and scale

Adaptive boosting classifiers (AdaBoost) machine learning algorithm was first formulated by Freund and Robert (1995). AdaBoost works with learning classifiers arranged into a cascaded structure. Each classifier in the cascade consists of a linear combination of weak classifiers. Each weak classifier learns only one simple feature. The process of boosting proceeds as follows. AdaBoost calls a weak classifier repeatedly in a series of rounds. For each round, a distribution of weights in the classifier is updated. This process indicates the importance of examples in the data set for the classification. In each round, the weights of each incorrectly classified example are increased (or alternatively, the weights of each correctly classified example are decreased). This way the classifier learns to reject as many non-faces as possible and retain only faces. A new classifier at a higher level of the cascade focuses more on the misclassified examples. This structure is adaptive in the sense that each subsequent classifier in a cascade works with samples misclassified by previous classifiers.

Figure 2.18 illustrates a face detector based on a cascade of boosted classifiers proposed by Viola and Jones (2001; 2004). In this method, the facial image is represented by rectangle Haar-like features called "integral image" (Papageorgiou *et al.*, 1998). The classifiers are arranged in a cascade structure in which weak classifiers are followed by increasingly more complex classifiers. Adaboost classifiers select a small number of the

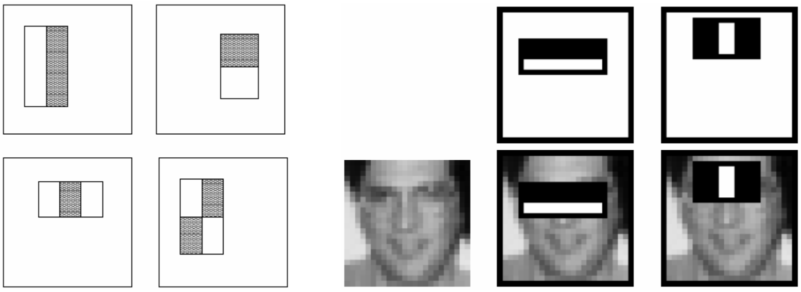


Figure 2.18. The left image shows examples of rectangle features. The sum of the pixels which lie within the light rectangles is subtracted from the sum of pixels in the dark rectangles. The right image shows rectangle features superimposed on a typical training face. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose. Reprinted with permission.[18]

[18] Reprinted from *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'01)*, 1, 2001, Viola P. and Jones M., "Rapid Object Detection Using a Boosted Cascade of Simple Features", Fig. 1 (p. 512) and Fig. 3 (p. 514), copyright © 2001 IEEE, with permission from IEEE.

critical rectangle features from the input image. The weak classifiers process a large amount of data and discard the background regions leaving more computation on promising face-like regions for complex classifiers. This allows faces to be detected fast and efficiently.

Adaboost can be also applied at a feature level for the detection of separate facial landmarks as in (Cristinacce & Cootes, 2003; 2006). Later, the extended set of Haar-like features was proposed for the detection of the face (Lienhart & Maydt, 2002) and facial features (Wilson & Fernandez, 2006; Lu *et al*., 2007). Vukadinovic and Pantic (2005) proposed performing GentleBoost on Gabor-based features allowing the detection of facial feature points from images of expressive faces.

The main advantage of AdaBoost classifiers is that they do not have a dimensionality reduction problem in contrast to PCA and NN. In the cascade of weak classifiers, each weak classifier works with a reduced amount of data from the preceding classifiers. This enables fast and efficient classification. AdaBoost learning strategy can work with other learning algorithms in order to improve their performance. AdaBoost is largely sensitive to noisy data and outliers. However, if it is trained on a large representative set of images it can achieve high performance. AdaBoost-based methods are mainly used for face and feature detection (Yang *et al*., 2002).

Advantages of Adaboost-based methods are that they:

- work with grey-scale images
- are adaptive while reducing the amount of information to be processed at subsequent stages of the cascade
- if trained on a representative image set, they are effective and efficient under many conditions
- are fast to apply in real-time

Disadvantages of Adaboost-based methods are that they:

- have a problem in the selection of training data and training procedure
- require many positive and negative examples
- are time consuming in training and testing
- need to search for face and facial landmarks over space and scale

2.4 Performance Evaluation

The output of the face and feature detection system has a wide range of applications. The found face and feature positions in static image or video can be delivered as an input to various systems of automatic face analysis such as face recognition, facial expression analysis, and perceptual vision-based HTI. In this section, the issues on a performance analysis of face and feature detection systems presented in the literature are discussed. In order to allow systems of automatic face analysis to work fully automatically, efficiently, and robustly in real time, it is important to achieve fully automatic, robust, accurate, and fast face and feature detection. In practice, the performance of the face and feature detection system is examined in terms of the form of the detection output, detection accuracy, detection rate, false alarm rates, and its speed. Let us consider these evaluation metrics in more detail.

Form of the Detection Output

The choice of a form of detection output is an important issue to be considered in the performance evaluation of the face and feature detection systems. The form of the face and feature detection output data is directly related to its detection accuracy and is driven by the application for which the detection system has been designed. It has been shown that the majority of face analysis systems require a precise localization of eye centres. Eye centres are widely utilized, for example, in face normalization and alignment, which are typically done prior to PCA-based face recognition. The eye centres have to be detected with high accuracy because these methods are sensitive to displacement and scaling errors in face detection (Zhao *et al*., 2003). Other face recognition systems require a large number of precisely located facial features from the eye, nose, mouth, and sometimes cheek or chin regions (Wiskott *et al*., 1997; Colbry *et al*., 2005; Campadelli *et al*., 2007). Still other face analysis systems take a facial region (like that demonstrated in Figure 2.19) rather than a feature point as their input (Zhao & Pietikäinen, 2007). Because of this, it was recommended (Campadelli *et al*., 2007) that a face and feature detection system should clearly specify the form of its detection output and the range of applications for which this output is suitable. At the same time, it was recommended that face analysis systems should define their input data requirements in terms of data type (*i.e.* detected features) and required accuracy of the detection result.

Generally, the following widely utilized face and feature detection data outputs can be defined. Eyes are usually defined by eye centres, eye corners, exact eye contour, or bounding box which contains the eye region. If eyebrows are chosen to be detected, their position is

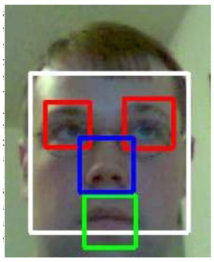


Figure 2.19. Example of the FFFD detection output data in a form of rectangular boxes placed over the found facial regions (Wilson & Fernandez 2006). Reprinted with permission.[19]

[19] Reprinted from *J. Computing Sciences in Colleges*, *21 (4)*, 2006, Wilson Ph.I. and Fernandez J., "Facial Feature Detection Using Haar Classifiers", Fig. 5 (p. 135), copyright © 2006 Consortium for Computing Sciences in Colleges (CCSC), with permission from CCSC.

usually evaluated by a bounding box which includes the eyebrow region, exact contour curvature, or a number of points on the eyebrow. Nose is defined by centre of the nose tip, centre coordinates of nostrils, or a bounding box which contains the nose region. Mouth is defined by the mouth centre point, mouth corners, exact mouth contour, or bounding box which contains the mouth region. The accuracy of detection of the facial pattern as a whole is usually defined by the locations of the eye centres or bounding box containing a face region.

Detection Accuracy

The *detection accuracy* of face and feature detectors with output in the form of a point measure is given in terms of either visual inspection of the detection result or allowable error range. Visual inspection cannot be considered an appropriate evaluation measure as it is a subjective decision and cannot give objective criteria of what to consider a correct detection result. The *error range* is usually calculated as a distance between manually annotated and automatically detected feature point locations. Allowable error range has usually been reported in terms of Euclidean pixel distance - the smaller the number of the pixels, the better is the accuracy of the detector.

According to a point evaluation measure proposed by Jesorsky *et al.* (2002), a detection result is considered correct if the distance between manually annotated and automatically detected feature point location is less than 1/4 of the annotated interocular distance[20]. This measure is normalized to the face scale and, therefore, ensures the possibility to compare the detection results from different studies and databases. This measure has already been widely adopted by many studies for purposes of assessing the quality of eye localization (Ma *et al.*, 2004; Tang *et al.*, 2005; Niu *et al.*, 2006).

Recently Rodriguez *et al.* (2006) elaborated this issue further and proposed new face and feature detection accuracy measures. New measures appeared to be more discriminative than the former one as they permit a quantitative evaluation of the face recognition degradation with respect to different error types. Instead of considering only the Euclidean distance between the detections and the annotated points, the new measure considers four types of errors: horizontal and vertical displacements, scale errors, and rotation errors. Studies on deformable template-based detection methods output contour of the faces and landmarks detected. These methods typically utilize the point evaluation measures described above to report their findings (*e.g.* Campadelli *et al.*, 2007). It is generally unclear how the accuracy of the contour detection can be defined in this case.

[20] The interocular distance is the distance between the centres of the eyes (Farkas, 1994).

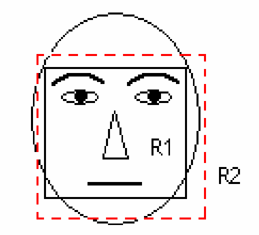


Figure 2.20. The detected area for manually annotated face (solid line) is R1 and the detected area for automatically detected face (dashed line) is R2.

The measures described of face and feature detection accuracy are point measures sufficient for those applications which can make use of a single pixel point result as an output of face and feature detectors. However, there are applications requiring a detector to find a region in the face rather than to give a point solution. Studies on this topic have proposed several observations to define a detection result as a correct one. For example, a proportion index was calculated that showed a relative proportion between manually annotated and automatically detected face or feature areas (*e.g.* R1 and R2 from Figure 2.20). If the proportion index is smaller than a predefined threshold, the detection result is considered to be correct. Other studies checked if the centres of the chosen features lay inside a manually annotated bounding box defining the boundaries of the face or its features. For example, both eyes are inside a box which bounds the face or the centre of the landmark lies inside a box which bounds the landmark. Some studies on edge-based landmark localization (*e.g.* Golovan *et al.*, 2001) calculated coefficients of selectivity as a ratio between the total number of edge points found inside the feature region and a general number of edge points extracted from the image.

Detection Rate and False Alarms

The *face detection rate* is usually calculated as the ratio of the number of correctly located faces to the total number of faces used in testing. The *false alarm rates* (*i.e.* false negative and false positive rates) of face detection are usually given as the ratio of the number of incorrectly detected faces to the total number of faces used in testing. *False negative* shows how many faces were rejected as not belonging to face class during the test. *False positive* shows how many non-face regions were classified as faces during the test. Similarly, the *feature detection rate* is usually calculated as a ratio of the number of correctly located features to the total number of features used in testing. The *false alarm rates* of feature detection are usually computed as the ratio of the number of incorrectly detected features to the total number of features used in testing.

Speed

The execution time of the face and feature detection system is an evaluation metric which is usually applied for static input data. It shows how much time a system requires to perform an analysis of a single static image. The *frame rate* of a face and feature detection system shows how many images (*i.e.* frames) of a given size the system is able to process within a certain amount of time. The frame rate is usually measured in frames per second (FPS). Generally the speed requirement for a real-time face and feature detection system is about 20-24 FPS. This speed is sufficient for the majority of real-time face analysis applications (Turk and Kölsch, 2004).

2.5 APPLICATIONS

Automatic face and feature detection is an essential initialization step in all systems of video-based automatic face analysis. In this section, three important applications of face and feature detection output data are described. These are face recognition, facial expression analysis, and perceptual vision-based HTI. Figure 2.21 illustrates a generally widely accepted way of how facial data flow is processed in these systems. After the facial data signal is acquired by an imaging device in the form of a static image or video it is delivered to the face and feature detection system. Some preprocessing can be done prior to this step, for example, face segmentation. Further, face and feature detection data are transferred to the following stages of face analysis. Before that, the data can be optionally preprocessed, for example, a face can be normalized according to its scale or pose.

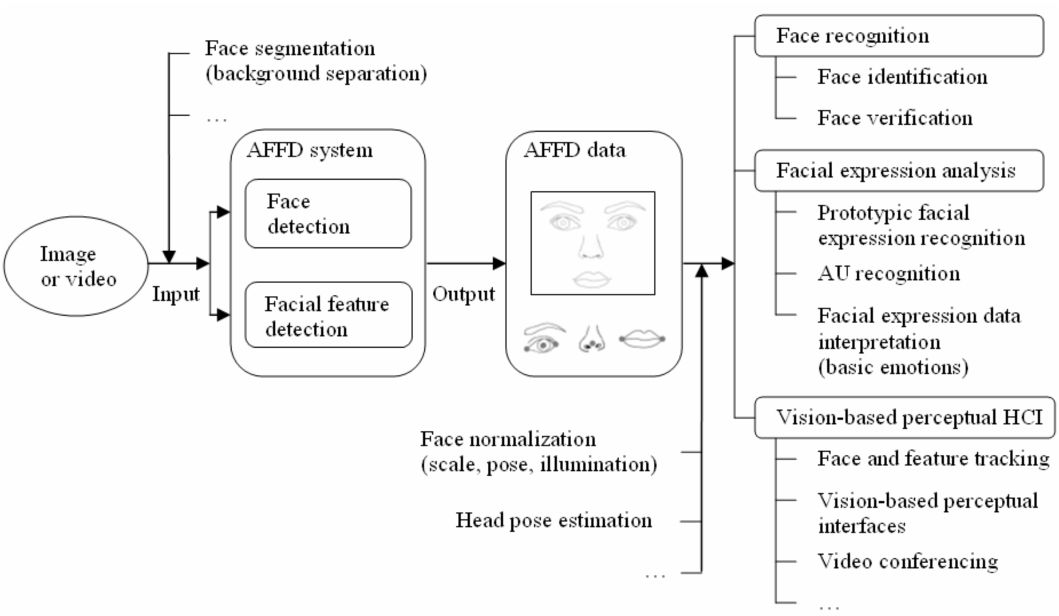


Figure 2.21. Multiple application fields of automatic face and feature detection (AFFD).

Face Recognition

Face recognition is a method of biometrics for identifying individuals by the features of the face. It can be face identification or face verification. *Face identification* is defined as a process of classification of the input face to one of the existing face classes stored in the database or rejection of the input face as an unrecognized/unknown face. *Face verification* is defined as a process of confirmation that the input face has an established identity or rejection of the input facial identity. Many systems of face recognition rely on the detected eye locations (Burl & Perona, 1996; Zhang *et al*., 2005). Some other face recognition techniques may require more features than just the eyes (Heisele *et al*. 2003; Colbry *et al*., 2005; Campadelli *et al*., 2007). Four points - the eyes, nose and mouth are required by all face recognition systems derived from subspace methods in order to warp a face region before projection (Shakhnarovich & Moghaddam, 2004). Other techniques operate on larger sets of facial features. For example, Wiskott *et al.* (1999) based their recognition process on the local processing of image texture in the neighbourhood of several characteristic points. Having such variability in face recognition schemes and ways of their initialization, it is important to achieve compatibility between face and feature detection data output and face recognition input data requirements.

In theory, face recognition algorithms rely heavily on the data output from face and feature detection, especially if they require face normalization or alignment (Zhang *et al*., 2005). It has been demonstrated that face recognition suffers greatly from imprecise localization of the face components (Campadelli *et al*., 2007). PCA-based holistic face recognition has been shown to be particularly sensitive to displacement and resolution errors (Turk & Pentland, 1991; Burl & Perona, 1996; Zhang *et al*., 2005). In practice, however, face recognition systems usually avoid dealing with the issue of face and feature detection and do not clarify how they detect facial components for face recognition. Many face recognition systems skip face and feature detection and assume perfect localization of face and facial features by relying on manual annotations of facial components in the image or in the first frame of the video sequence.

Facial Expression Analysis

Automatic facial expression analysis has attracted much attention in recent years. Pantic and Rothkrantz (2000*a*) introduced an exhaustive review on the methods of facial expression analysis and related problems. They stated that "in the case of static images, the process of extracting the facial expression information is referred to as localizing the face and its features in the scene." In other words, as in face recognition, facial expression analysis schemes rely heavily on face and feature detection data output.

In facial expression recognition systems, after the face or feature has been detected, the next step is to extract and represent facial changes caused by facial expressions. The systems of facial expression analysis mainly operate on the information contained in the region of prominent facial landmarks. Therefore the most frequently cited features to be used in facial expression analysis are prominent facial landmarks which are eyes, eyebrows, nose, and mouth. For example, a smiling expression is characterized by stretching the mouth to the sides and laterally upwards and wrinkles appearing in the outer corner of the eye. Anger can be detected from the presence of upper and lower eye lid tightening among other facial changes. Disgust is characterized by the appearance of wrinkles in the area of the bridge of the nose and a specific mouth shape that resembles an upside down v-curve (Ekman & Friesen, 1978). Other facial expressions may have important clues in the appearance of secondary facial landmarks located at a distance from prominent facial landmarks. For example, surprise often produces wrinkles in the forehead, and frustration might change the appearance of chin. Still, in order to detect these expressive changes, the prominent facial landmarks are usually detected first. They serve as "stable" facial features which guide the detection of the secondary facial landmarks.

It has been shown that even though much progress has been made in this field, the task of expression recognition with high accuracy remains difficult due to the complexity and variety of facial expressions. Since the publication of Pantic and Rothkrantz's survey (2000*a*), further progress has been achieved to date in the field of facial expression analysis. Many authors (Lien *et al.*, 2000; Pantic & Rothkrantz, 2000*a*; Tian *et al.*, 2002; Michel & Kaliouby, 2003; Feng *et al.*, 2005; Zhao & Pietikäinen, 2007) have develop their own face and feature detectors, while some still initialize their facial expression recognition techniques on manually annotated feature positions. One important step in improving existing facial expression recognition systems is to achieve a fully automatic, robust, efficient, and fast face and feature detection.

Vision-Based Perceptual HTI

Another possible application of face and feature detection data relates to the quality of vision-based perceptual HTI. There are two ways to utilize facial information in HTI which correspond to voluntary produced and spontaneously expressed visual changes in the face. The former is to use facial information as a mean of voluntary input to the computer in control situations, when the user consciously produces movements and gestures to control the application. These behaviours and movements are made voluntarily and do not necessarily reflect the affective state (*i.e.* cognitive or emotional experience) of the user. The control situation may, for example, by an eye tracking system in which eye movements are used for typing and pointing in graphical interfaces (Majaranta & Räihä, 2007). Another example is a video game controlled by gaze direction. In a gaze-based unimodal control application, the gaze control includes commands which allow navigation in the game universe, initialization or termination of movements of the game characters and objects, and changes in the avatar's appearance. Recently, a multimodal approach has also been proposed to create "face interface" that utilizes facial expressions along with gaze direction, both produced on a voluntary basis, for the purpose of application control (Surakka *et al.*, 2004). In this control situation, gaze direction serves as a mean of pointing and bioelectric signals resulting from facial movement activations serve as means of selection of objects in graphical interfaces. A non-invasive visual-based "face interface" can potentially incorporate facial expressions, gaze direction, head movements - all analysed automatically, robustly, and efficiently in real time.

The second approach makes use of the user's spontaneous facial behaviours as a means of involuntary input from user to computer in order to allow the application to adapt to the user's needs. There is a lot of evidence that people do respond socially and emotionally to the computer through many modalities despite the fact that they know that computer is not yet meant to have a social intelligence (Reeves & Nass, 1996; Turk, 2005; Picard, 2002; Partala *et al.*, 2006; Surakka and Vanhala, 2008). Thus user models developed to enhance the understanding of the intended meaning of the user's behaviour should take into account both the affective information and the possible emotional response of the user. A facts growing research area called affective computing (Picard, 1997) is concerned with the design of systems and devices which can recognize, interpret, and process socially and emotionally associated facial signals in HTI. The main idea of affective computing is that if a computer could recognize a user's emotion, it could better adjust the computer's behaviour to the user's behaviour. Affective computing requires a solution to two main problems - searching for appropriate emotional signals that might be applied to HCI on the one hand and solving technical difficulties involved in the detection, recognition and interpretation of the user's emotional signals on the other.

It should be mentioned that there is no direct correspondence between facial expressions and human emotions. The recognition of human emotions requires knowledge from multiple modalities including visual, audio, and physiological signals considered together with the context of the situation (Anttonen and Surakka, 2005; Partala *et al.*, 2006). Although facial expressions are not necessarily complete reflections of spontaneous emotions, they are frequently associated, to some extent, with emotional categories. Emotion recognition technology is currently being developed in order to detect, recognize and interpret the emotional signals received from the computer user. The modern level of technological progress enables receiving emotional signals with relatively low signal-to-noise level. But other technical difficulties concerning the evaluation of emotional signals show that the design of more complex computational algorithms still has a long way to go.

2.6 Summary

Automatic face and feature detection involves two important aspects - selection of a face representation and design of a feature detector. The main requirement is that the face representation characterizes the face or facial parts in a robust, distinctive, compact, and easy-to-compute way. It was demonstrated that face representation is dependent on the nature of the image data being processed (*e.g.* colour or grey-scale image, static image or video sequence) and the application at hand. It was shown that careful selection of the appropriate face representation is of great importance because a feature detector will fail to achieve accurate detection results if inadequate features are used. It was also shown that the design of a feature detector may vary from application to application depending on the face representation used.

Two different face representations were introduced – local and global face representations. In some cases, the use of a global representation has its own advantages. For example, preliminary face localization enables fast feature extraction because it discards those parts of the image that are irrelevant for the feature search. However, global face representation has serious limitations. First, because the face is a highly non-rigid visual object, global representation is highly sensitive to occlusions and changes in facial expressions, head rotations, and lighting. For example, in a presence of between- and within-individual variations of facial expressions, it is difficult to enumerate all possible face models without a hierarchical decomposition of the problem. In addition, global representation requires processing of high dimensional data, which is a difficult problem.

In contrast to global face representation, appearances of single facial landmarks exhibit less variability. Facial features analysed locally usually vary less in pose, lighting, and facial expression changes than the pattern of the whole face. Therefore, by focusing first on facial parts, a local image representation makes it possible to handle a wider range of conditions than the global one. Generally, it has been shown that local face and feature detectors outperform global ones in a number of conditions including head rotations, changes in lighting, occlusions, and facial expressions (Yow & Cipolla, 1997; Pantic & Rothkrantz, 2000*a*; Heisele *et al.*, 2001; 2006; Tong *et al.*, 2007). The advantage of this approach is that it is robust in a wide range of viewpoints, even at profile views. Local face and feature detectors have been shown to achieve better robustness regardless of facial expressions (Pantic & Rothkrantz, 2000*b*; Campadelli *et al.*, 2007). Utilizing a bottom-up approach, local detectors operate in a small search space by reducing false or impossible feature constellations in the early stages of processing, resulting in efficient and fast computation. Partial occlusions in the face (*e.g.* by hands, hair, or eye-glasses) as well as self-occlusions resulting from facial expressions are serious constraints for face and feature detectors. Local detectors overcome these constraints because occlusions affect the outputs of only a few feature detectors at a time. For example, if nose detection fails it is still possible to restore the position of the mouth in the image on the basis of the eye and mouth locations found. Therefore, a solution to the occlusion problem is in the development of a feature detector that is robust against changes in a small number of its input features (Heisele at al., 2006).

However, there are methods using a combination of local and global face representations. In many cases, such hybrid methods give better results and can cope with different variations in facial appearance. For example, the method of Viola and Jones (2004) explained above detects faces with a set of simple classifiers operating on locally computed image features. Each of these simple classifiers is applied to a fixed coordinate position of the image preserving the holistic image representation. Another frequently used hybrid approach is the use of learning classifiers on local features, for example, wavelets. This approach provides good face and facial landmark detection results (Chen *et al.*, 2004; Fasel *et al.*, 2005; Vukadinovic & Pantic, 2005; Campadelli *et al.*, 2007). Hybrid methods also have the advantage of being biologically plausible as they imitate bottom-up and top-down mechanisms of face processing in the brain. Bottom-up face processing includes the extraction of local properties of a face like edges, colours, and intensities (Golovan *et al.*, 2000; 2001; Shaposhnikov *et al.*, 2002). Top-down face processing is guided by the knowledge of holistic facial properties, for example, knowledge of a typical human face (*i.e.* structural and symmetrical facial properties), behavioural clues, and knowledge of how faces might look in different conditions (Cristinacce & Cootes, 2006; Campadelli *et al.*, 2007).

Apart from a general classification of face and feature detection methods into global and local, another classification was presented which divided the detection methods into three significant categories, namely, feature-, template-, and learning-based methods. The pros and cons of the methods belonging to each of these approaches were analysed so as to be applied to the task of real-time, robust, and efficient face and feature detection. The robustness of each method category was considered in a large framework with regard to changes in lighting, head rotation, occlusions, and facial expressions. It should be noted that the given classification into three categories is somewhat conditional. In fact, only few methods can be classified solely into one of these categories.

There is no simple answer as to which method is the best one. In the survey, any comparisons of the published detection rates of the methods presented were deliberately omitted. Such attempts have been undertaken earlier in other reviews (Hjelmas & Low, 2001; Yang *et al.*, 2002; Zhao *et al.*, 2003). However, one should evaluate the results of these reviews with extreme caution because rough and possibly misleading comparisons can be made. There are several reasons for this. First, some studies used commonly available image databases, while others created their own databases which are not easily accessible for evaluation. Even if the same image dataset was used, there is no way to know the exact meaning of the detection rates as different preprocessing of the images can be used and different training sets can be chosen in algorithm development. It is known that the use of the same database over and over again helps to "tune" methods, particularly for this dataset. This enables better results for this particular database. In this case, a comparison between different face and feature detection methods can be made in favour of more "tuned" methods.

On the other hand, it is not usually clear which features are chosen for detection. Many authors do not specify this when reporting their results. For example, if the rates are reported for eye detection, a clear definition of what exactly was detected is required. Was it a region in the image? Was it the coordinates of eye centre or eye corners? Or is it a contour of the eye pupil that was detected? On the other hand, these issues appear less important given that after decades of research in face and feature detection there is no general consensus among researchers on the criteria for the evaluation of the detection results. This is especially true regarding those detectors which output region-based result. To the best of my knowledge, there are no precisely defined criteria for the evaluation of the landmark localization result represented by a region in the image, not a single point. This dissertation addresses this issue later in Chapter 4. A new rectangular measure of the landmark localization result will be introduced in the form of a rectangular box.

Nevertheless, a comparison can be made as some method categories generally proved to be more or less sensitive to facial expression variations in the facial appearance. In this respect, the advantages and shortcomings of global and local detection methods were already discussed. As to the robustness properties of the methods classified into feature-, template-, and learning-based categories, they can be generally summarized as follows. It can be said that with a few exceptions, feature-based face and feature detection methods have been seen to be generally affected by facial expressions. Edge-based, wavelet-based, and texture-based methods are all generally stated to be sensitive to facial expressions to some extent. The difficulty of evaluating the robustness property of these methods is that the effect of facial expressions has not been systematically studied and reported in the literature. Only occasionally do the authors mention specific facial behaviours which impaired a detection process (Lien *et al.*, 2000; Tian *et al.*, 2002). One can see the importance of such analysis as it helps to improve the existing methods and develop new ones which will be largely invariant particularly to expressive deformations in the face.

Additionally, a majority of the papers simply state that the databases they used included facial expressions. However, a description of the facial expressions tested is not always given. This can be accepted for public databases as it is always possible to check the variability and level of complexity of the expressions from these databases. However, it is not always possible to do so for those datasets created in closed labs and not available for public appraisal. One possible way to overcome this limitation would be to create a tradition of describing facial expressions while reporting scientific results on the performance of the detection methods on the expressive databases. The next chapter presents several options for expression classification in terms of prototypical facial displays and at the level of single muscle contractions.

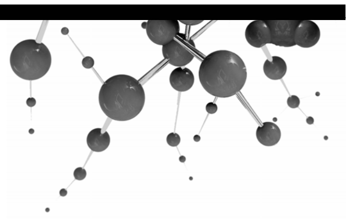
Invariance of the detection methods to facial expressions achieved at the feature level would mean a number of important advantages. Firstly, facial expression analysis could be already performed at the early stages of processing and utilized throughout all subsequent steps of the automatic face analysis. Secondly, a decrease of computation cost in run-time could be achieved by applying a set of expression-invariant feature-based detectors at all stages of feature-based face analysis. For example, the same features could be utilized for feature-based face detection and for other face analysis tasks like feature-based face recognition and feature-based facial expression analyses. Nowadays, however, face detection, face recognition, and expression analysis often utilize different features. Finally, it is obvious that the number of false detections could decrease if expression-invariant features are used. Altogether this will lead to a better performance of the detection methods resulting to better performance of the following steps of automatic face analysis.

In the case of template-based face and feature detection methods, it was demonstrated that predefined facial templates are unlikely to be successfully applied to the task of expression-invariant face and feature detection. Due to the rigid structure of the face, facial expressions have to be modelled by a great number of facial templates. The task of enumerating all possible templates is arduous and the expected performance is poor. The same can be applied to other affecting variables like, for example, head rotations and occlusions. On the other hand, there are a number of deformable templates which proved to be robust against facial expressions. It is a well established fact that snakes, for example, are able to track the shape of the mouth regardless of such mouth deformations as open and closed mouth, pursed lips, and visible teeth. As mentioned, the only limitation of this method is that snakes need to know in what part of the image the feature is located. In other words, snakes need to be initialized in a very close neighbourhood to the feature. This can be done by image preprocessing with feature-based methods, for example, colour-based face detection can be applied to define the approximate location of the face and followed by more precise detection of eyes and mouth by utilizing edge or intensity map analysis.

Starting at the beginning of the 2000s, a considerable shift in face and feature detection system development was made to the use of learning-based classifiers. Since then several high-speed efficient learning-based face and feature detectors have been proposed. It is possible to predict that the introduction of new and the improvement of existing learning-based schemes will continue the development of the face and feature detection field in the future. At present some of the existing face and feature detection learning-based classifiers may be computationally expensive and require time-consuming training and testing phases. The learning-based classifiers have to be trained (usually by a human operator) on large representative sets of face and non-face images. This entails numerous problems concerning the selection of the training set and performance of the training procedure. For example, it is generally acknowledged that it is difficult to collect a training set of images covering a spectrum of possible face variations caused by facial expressions, head pose variations, and occlusions. Hybrid face and feature detectors combining the learning-based high-level classifiers operating on local expression-invariant features could solve these problems.

More work is still required in the field of automatic face and feature detection. There is a constant search for more compact and robust face representations enabling the introduction of more robust, accurate, efficient, and fast face and feature detection schemes. The question of the development of expression-invariant methods is one of the tasks still in need of further investigation. There is an increasing need for real-time face and feature detection schemes in many application fields. Thus, another task is to optimise face and feature detection schemes to reduce the number of single computations and time required for processing one single image or video frame. In the future, as hardware and software development advance, it may become possible to optimise and speed up existing face and feature detection schemes which are not at present applicable in real time. Now, however, one needs to make a compromise between the extent of complexity in the computational models and processing algorithms and the desired speed of the detection systems. The understanding of the mechanism of biological vision on the other hand might yield important insights into the design and implementation of computer vision algorithms. The current state of development in computer vision already acknowledges a clear necessity for redundancy reduction in visual input fed to face and feature detection systems. In many systems, information compression is done similarly to the way it is done by biological visual systems in the early stages of face detection. Edge-based and wavelet-based representations of the face have demonstrated their usability in this case.

In conclusion it can be stated that face and feature detection is interesting, challenging, and useful for many applications research area. Great research efforts have been devoted to the task of automatic, efficient, and robust face and feature detection. The field is rather wide and not all approaches to automatic face and feature detection were described in this short survey. However, despite the increasing volume of related literature, face and feature detection is still an unresolved problem. The existing face and feature detection techniques need improvement in detection accuracy and robustness with regard to changes in facial appearance. Recently, the issue of expression-invariant face and feature detection has gained considerable attention as more and more facial expression databases become available to the research community. A number of interesting methods which can deal with expression variations in the face have been proposed. However, a more detailed and systematic approach to expression-invariant face and feature detection is still required.



3 Facial Expressions

In this chapter, issues related to facial expressions and their automatic analysis are discussed in more detail. Achievements in the analysis of facial behaviours which apparently influenced onto the progress in the automatic analysis of expressive faces (AAEF) are considered. On the whole, the purpose of this chapter is to demonstrate a successful knowledge transfer made from behavioural science to computer vision research.

From birth, humans learn to perceive and display facial expressions. In spite of between- and within-individual variations in facial behaviours, adults generally demonstrate an astonishing ability to recognize facial expressions in an effortless manner. Given the importance of facial expressions and their presence in our everyday life, it is no wonder that humans aimed at making computers able to "read" expressive faces automatically. Broadly speaking, automatic analysis of expressive faces refers to all systems of computer vision which attempt to automatically analyse faces modified by facial expressions in static images and video sequences. The ability of the computer to process faces invariantly with respect to facial expressions would increase the efficiency of the systems of automatic face analysis and, as a result of these, face identification and verification. On the other hand, the ability of the computer to detect expressive changes in the face and interpret them into meaningful signals would give rise to new ways of social-emotional HTI (*e.g.* facial expressions as pointing devices) and more profound analysis of the user's emotional and cognitive states. As already pointed out, there is no direct correspondence between AAEF and automatic analysis of human emotions. The analysis of human emotions deals with a more general problem and requires knowledge from multiple modalities including visual, audio, and psycho-physiological signals considered together with the context of the situation (Anttonen & Surakka, 2005; Partala *et al.*, 2006).

Significant progress has been achieved in AAEF, especially during the past few decades. However, unlike human abilities in processing expressive faces, computer-based AAEF systems do not yet cope with a whole range of variations in facial appearance brought about by facial expressions. Looking back at the history of progress in AAEF, it can be noticed that even though the first AAEF algorithms appeared at the early stages of the development of computer-based face analysis techniques, the progress in this field was rather slow. The one reason for this was the fact that for a relatively long time in the AAEF domain there were neither commonly accepted standards of expression categorization nor publicly accessible databases of labelled facial expressions.

The first systems of automatic face analysis appeared in the early 1970s (Sakai *et al.* 1971; Kohonen, 1977) - as soon as computers were able to process large amounts of data (*i.e.* digitized photos). After that for several decades face recognition, for example, has been applied to a scenario in which the person to be recognized was relatively expressionless. At that time, few studies addressing expression-invariant face recognition set the problem as a general "neutral vs. expressive" face recognition problem (Yacoob *et al.*, 1995). Because of this fact, there was no clear evidence about how facial expressions attenuate the recognition process. In other words, the results reported for example, for the "smile" database did not state if it was a smile with open or closed mouth, if there were other facial movements involved, and what the intensity level of the expression was.

Later, expression classification in AAEF fell into three categories representing expressions according to its structure, emotional context, and intensity. The classification made according to structural changes in the face referred to the movements of prominent facial features such as "raised eyebrows", "mouth corners down", "smile", and "frowning" (Black & Yacoob 1997). Another classification of expressive faces was made with regard to the emotional state of the person in the image, for example, "happy" or "angry" (Zhang *et al.*, 1998). The classification of the expression on the base of its intensity was made in terms of "neutral", "weak", and "strong" expressions.

This way, the level of expression categorization adopted by AAEF researchers at that time was "imprecise, ignoring differences between a variety of different muscular actions to which they may refer, and mixing description with inferences about meaning or the message which they may convey" (Hager & Ekman, 1995). This way, the lack of standards of expression categorization put obstacles in the way of AAEF by the lack of a detailed and widely appreciated description of face deformation categories. Consequently it was difficult or even impossible to evaluate and compare results from different studies dealing with automatic face analysis. Whereas AAEF researchers lacked deep knowledge of human facial behaviour, scientists from other disciplines had long been interested in the topic of facial expressions. Since the work of Darwin (1872), intensive studies in psychology, behavioural sciences, and physiology had been conducted to accumulate solid empirical evidence and good theoretical basis for analysing how different facial muscle activations change the appearance of a face during emotional, social, and cognitive reactions. This knowledge of facial expressions was little by little integrated into computer science and now computer vision researchers can benefit from it.

The next section presents a short introduction to the nature of facial expressions. This is followed by a description of two main approaches to facial expression categorization originating from behavioural science research. Both expression categorizations are nowadays largely accepted in the field of computer vision. Next, different existing facial expression databases are considered from the point of view of their applicability to AAEF research. Facial expression data are viewed as useful tools for the development and testing of face and feature detection methods. The four expressive databases used in this dissertation are considered in more detail. The chapter ends with a short summary.

3.1 INTRODUCTION TO FACIAL EXPRESSIONS

Facial expressions are emotionally, socially, and otherwise meaningful reflective signals in the face. Facial expressions play a critical role in human life providing an important channel of nonverbal communication. Facial expressions have several communicative functions. For example, they are the primary channels of conveying emotional information to others. This way, facial expressions develop and regulate interpersonal relationships. Co-occurring with a verbal message, facial expressions coordinate the flow of conversation while, for example, regulating the exchange of speaking turns, understanding speech articulation, and emphasizing what has been said. (Ekman, 1982)

Anatomically, facial expressions result from contractions and relaxations of different facial muscles. “These [non-rigid] movements [...] pull the skin and tissues, temporarily distorting the shape of the eyes, brows, and lips, and the appearance of folds, furrows and bulges in different patches of skin.” (Hager & Ekman, 1995). Facial expressions result not only in considerable changes of feature shapes but also in changes in their location in the face, out-of-plan changes (*e.g.* showing the tongue), and self-occlusions (*e.g.* closed eyes, bitted lips, and visibility of teeth or tongue).

Facial expressions drastically change the appearance of permanent and transient facial features (Rinn, 1984). Transient features are wrinkles and protrusions resulting from momentary deformations of soft facial tissues. Permanent facial features are facial landmarks such as eyes, eyebrows, nose, and mouth. Wrinkles can also be permanent features if they remain in the face when a neutral expression is displayed. Such wrinkles can be caused by age-specific modifications of the face or habitual facial expressions like frowning, which can produce permanently etched vertical wrinkles between the eyebrows.

Facial expressions of different structure rely upon differences in the proportion of muscle contraction and relative position of facial features (Duchenne de Boulogne, 1862). Apart from structural differences, facial expressions are also characterized by their intensity value and temporal course (*i.e.* dynamics). The intensity value defines the amplitude or the "strength" of the expression. The temporal course of an expression typically has three phases: onset, apex, and offset (Ekman 1979). Starting from neutral face, onset is the time when the expression appears and reaches its maximal intensity value. The apex is the time during which expression remains at its maximum intensity. The offset is the time when expression disappears and returns back to a neutral expression.

Facial expressions vary in their appearance both across human population and within the facial behaviour of a given individual, which can display expressions in different ways depending, for example, on the context of a situation. Facial expressions can be produced on a voluntary and involuntary basis. It has been shown that in some social situations humans adopt facial expressions on a voluntary basis (*i.e.* Duchenne and non-Duchenne smile (Surakka & Hietanen, 1998)). However, it was claimed (Ekman and Friesen, 1976; Ekman, 1982; 1984) that because expressions are closely tied to emotions, they are often involuntary. Thus, facial expressions are often considered as a reflection of the person's affective state or emotional experience.

3.2 Emotion-Based Expression Categorization

Although facial expressions are not necessarily complete reflections of spontaneous human emotions, they are frequently associated, to some extent, with emotional categories. There are two approaches to the emotion-based categorization of facial expressions - discrete and fuzzy approaches. The *discrete emotion-based approach* was established by Ekman and Friesen (1976) and implies the existence of prototypical human emotions. Prototypical emotions were postulated to be universal emotions across cultures and basic emotions as they evolved to help humans in dealing with fundamental life tasks. Ekman and Friesen (1976) argued that prototypical emotions are indeed discrete and form emotion families that differ from each other and other affective states (*i.e.* mood and trait) by a set of distinctive characteristics. The distinctive characteristics include prototypical signals of nonverbal communication, for example, facial expressions and body movements, and also physiology, appraisal mechanisms, and antecedent events. Ekman (1982, 1989) further suggested that prototypical facial expressions inherit the discrete nature of basic emotions and can be perceived as discrete entities with sharp category boundaries between them. He suggested that prototypical facial expressions can be assigned reliably to seven basic emotional categories: neutral, happiness, sadness, anger, fear, surprise, and disgust. In Figure 3.1, muscle contractions produce changes in the direction and magnitude of the skin surface motion resulting in the appearance of transient facial features which are characteristic for a particular prototypical facial display.

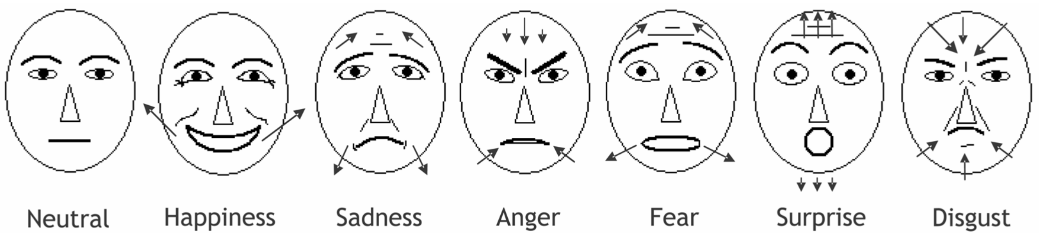


Figure 3.1. Prototypical facial expressions of emotions proposed by Ekman (1976) and primary directional cues of muscle displacements based on the suggestions of Bassili (1979).

The discrete approach implies knowledge of the prototypical emotional categories to which the expression can be assigned. In real life, however, "pure" prototypical expressions occur relatively infrequently. This means that not all facial behaviours can be directly classified under seven basic emotional prototypes. Instead, very often emotion is expressed as multiple simultaneous facial expressions.

This is done by the *fuzzy emotion-based approach* to facial expression categorization (Russell & Bullock, 1986). This approach originated from the idea of a continuous space of facial expressions. According to this theory, facial expressions are fuzzy sets. Expression categories overlap each other with peak levels which correspond to prototypical facial displays. In this framework, Russel and Bullock (1986) proposed a linear structural theory of emotion concept with two dimensions - pleasure and arousal. This concept was then supplemented with idea that happiness and sadness are the most prominent bipolar dimensions in the linear expression space, followed by other dimensions. More recently, the idea of expression manifold appeared in AAEF providing a global, analytical representation of all possible facial expressions as manifold in a high-dimensional space with neutral expression as the central reference point (Chang *et al.*, 2004; 2007).

The debate about the universality of the emotional prototypes (Ekman, 1982; Izard, 1971) and their relation to the actual emotional state of the person (Ekman, 1989; Fridlund, 1991; Russell, 1994) continues. Increasing evidence of categorical perception of facial expressions has appeared in the literature (Beale & Keil, 1995; Calder *et al.*, 1996). Evidence recently emerged supporting both discrete and fuzzy expression categorization theories (Dailey *et al.*, 2002).

Many existing AAEF systems rely on the emotion-based categorization of facial expressions and classify the expressions examined into one of the basic emotional categories. However, emotion-based expression categorization has a number of important limitations. The first limitation is related to the intensity factor of the expression. As mentioned above, one of the main characteristics of facial expression is its intensity value. A particular facial expression can have different intensity values depending on the situation and the person displaying the expression. For example, facial expressions in everyday life communication are often displayed by small changes in facial appearance. A slight movement of one prominent facial feature can define the emotional reaction. For example, anger can be expressed by lips tightening and surprise can be displayed by slightly raising the eyebrow(s). However, the labels of prototypical facial displays lack information on the intensity level of the expression.

The next limitation of emotion-based expression categorization is that it does not consider between- and within-subject differences in the structure of the same facial expressions. For example, there are many ways to produce a smile. The main characteristics of an involuntary smile in happiness are rising of the lip corners obliquely upwards, narrowing of the eyelids, wrinkling at the corners of the eyes, and raising of the upper cheeks. By contrast, the "Pan American smile", that typically serves to show politeness at a voluntary basis, does not involve muscle activity surrounding the eyes. Additionally, a laughing face may also evoke such supplementary muscle activations as lateral stretching of the lips, tensing of the neck, and raising of the eyebrows. These variations in the facial appearance are lost in the much broader definition of the prototypical expression of happiness.

As AAEF technology developed, it has become possible to detect and track more and more detailed information about facial expressions. Computer vision researchers became interested in the analyses of tiny changes in the face brought about by different muscle activations. For these small changes in the face, emotion-based expression categorization was not descriptive enough. Thus, there was a need for a more detailed approach to expression categorization that would give a systematic description of facial expressions taking into account their intensity factor and structural differences.

3.3 MUSCLE-BASED EXPRESSION CATEGORIZATION

Facial Action Coding System (FACS) was developed for analysing all visually observable changes in the face brought about by activation of different facial muscles (Ekman & Friesen, 1978; Ekman *et al.*, 2002). FACS makes a distinction between facial expressions and emotional signals by classifying facial movements on a muscle-by-muscle basis rather than bringing them into one or multiple emotional categories. In other words, FACS represents an expression by coding facial muscle activities fully objectively, without referring to the emotional, social, or cognitive state of the person in the image. FACS systematically describes visible changes in the face as a result of single and joint muscle contractions in terms of action units (AUs). There are 46 AUs which code visible changes in the face produced by underlying facial muscle activations, and 12 AUs which describe changes in gaze direction and head orientation in coarser terms. The systematic coding approach provided by FACS allows for the decomposition and enumeration of the space of all facial expressions. The effect of more than one muscle activity can be combined into a single AU, for example, in the case of lowering the eyebrows and drawing the eyebrows together three muscles are involved and combined into one specific AU. There is the option to score the intensity of the expression on a five-level scale. Table I shows lower and upper face AUs and their short descriptive names (Ekman & Friesen 1978; Ekman *et al.* 2002).

Table I. Upper and lower face AUs descriptors from the FACS manual (Ekman et al. 2002).

Upper face AUs	Lower face AUs			
	Up-Down AUs	Orbital AUs	Oblique AUs	Horizontal AUs
1 Inner brow raiser	9 Nose wrinkler	18 Lip pucker	11 Nasolabial furrow deepener	14 Dimpler
2 Outer brow raiser	10 Upper lip raiser	22 Lip funneler	12 Lip corner puller	20 Lip stretcher
4 Brow lowerer	15 Lip corner depressor	23 Lip tightener	13 Sharp lip puller	
5 Upper lid raiser	16 Lower lip depressor	24 Lip presser		
6 Cheek raiser and lid compressor	17 Chin raiser	28 Lip suck		
7 Lid tightener	25 Lips part			
43 Eye closure	26 Jaw drop			
45 Blink	27 Mouth stretch			
46 Wink				

Prototypical facial expressions can also be analysed within the framework of facial muscles. No complete set of AUs corresponding to each prototypical emotion has so far been identified, although there is consensus about the key muscular actions involved in the emotional prototypes. Thus, certain specific combinations of AUs have frequently been suggested to represent six prototypical facial displays (Ekman *et al.*, 2002). Ekman and Friesen (1980) presented a database called Facial Action Coding System Affect Interpretation Dictionary[21] (FACSAID) that allows translating emotion related FACS codes into affective meanings. Figure 3.2 demonstrates examples of the correspondence between prototypical facial expressions and AU combinations.

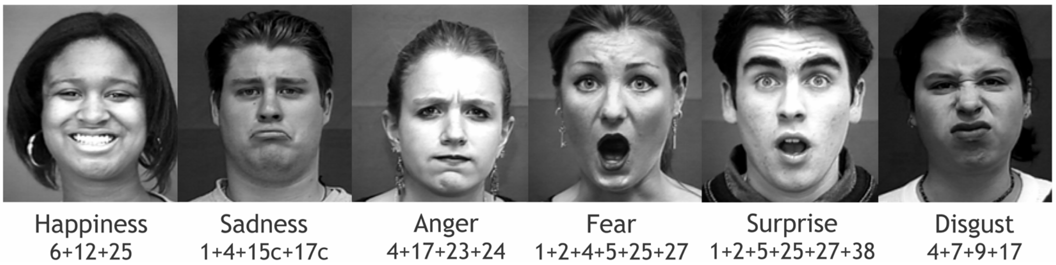


Figure 3.2. Prototypic facial expressions of emotions and corresponding AU codes. Images are from the Cohn-Kanade AU-Coded Facial Expression database (Kanade *et al.*, 2000). Reprinted with permission.

Nowadays, FACS is the most popular standard used to systematically categorize facial expressions. "To date, more than 300 people in all areas of the world have learned FACS and achieved inter-coder agreement on this test of proficiency, and many others have some degree of familiarity with this method, which has become a de facto standard for social, behavioural, and computer scientists studying the face" (Hager & Ekman, 1995). The activity in AAEF has greatly increased in recent years after the revelation of the FACS system and the appearance of large and publicly available databases of expressive faces. Nowadays, there is an emerging trend to design and train AAEF systems on AU-coded databases of expressive faces rather than emotion-based prototypical facial expressions (Pantic & Rothkrantz, 2000*a*).

It is fair to mention other existing schemes used to code facial expressions. Among others, there are the Emotional Facial Action Coding System (EMFACS) (Friesen & Ekman, 1983), the Maximally Discriminative Facial Movement Coding System (MAX) (Izard, 1979), the System for Identifying Affect Expressions (AFFEX) (Izard *et al.*, 1983). The listed coding schemes are only directed towards emotions. The object-based multi-media compression standard: Synthetic Natural Hybrid Coding (MPEG-4 SNHC)

[21] face-and-emotion.com/dataface/facsaid/description.jsp, (last accessed 26 June, 2008).

(Koenen, 2000) is a standard that encompasses analysis, coding (Tsapatsoulis *et al.*, 2000), and animation of faces (*e.g.* talking heads) (Hoch *et al.*, 1994).

3.4 Facial Expression Data

In the publications which follow this summary, the results on the performance of the developed methods were reported on four databases of systematically varied facial expressions. The databases include images of different structure and intensity. Three databases are publicly available and consist of static facial images showing expressions coded in terms of prototypical facial displays and AUs. The fourth database consists of short videos of prototypical expressions and was created in the laboratory of the Research Group for Emotions, Sociality, and Computing, Department of Computer Sciences, University of Tampere. A detailed description of each database is given below.

The Cohn-Kanade AU-Coded Facial Expression (Cohn-Kanade) database (Kanade *et al.*, 2000) is one of the most comprehensive collections of expressive images available to date (examples of images from this database can be seen in Figure 3.2). The database consists of image sequences taken from 97 subjects (65% female) of different skin colour (*i.e.* 81% Caucasian, 13% African-American, and 6% Asian or Latino ancestry) and ages varying from 18 to 30 years. Each image sequence starts with a neutral frame and ends up with an expressive frame labelled in terms of AUs. In the Cohn-Kanade database, AUs occur alone or in combination with other AUs. The images of the rest of the two static databases are labelled in terms of neutral and six prototypical facial expressions, namely happiness, sadness, fear, anger, disgust, and surprise (Ekman, 1999). The Pictures of Facial Affect (POFA) database (Ekman & Friesen, 1976) consists of 14 neutral and 96 expressive images of 14 Caucasian individuals (57% female). On average, there are 16 images per facial expression.

The Japanese Female Facial Expression (JAFFE) database (Lyons *et al.* 1998) consists of 30 neutral and 176 expressive images of 10 Japanese females. There are about 30 images per facial expression on average. In the POFA and JAFFE databases, a particular expression varies in its intensity and structural configuration. The video database was created for the purpose of method testing under conditions of varying expression, out-of-plane head rotation, and real-time processing. The database consists of neutral, frowning, and smiling faces under three controlled head rotations with angles of rotation 0°, 20°, and 30° to both right and left. A low-cost Canon Mini-DV camera with 720x568 pixel image resolution and 24-bit precision for colour values was used.

In all databases used, the facial images display the face-and-shoulder region. The potential impact of lighting, background, large head rotations, and presence of facial hair or eye-glasses was controlled to some extent and therefore ignored. However, the effect of other affecting variables like ethnicity (*i.e.* skin colour), facial expressions, small out-of-plane head rotations, and presence of clothing, hair, and decorations was analysed. All the databases used include images with posed facial expressions, meaning that expressions were produced on a voluntary basis. The posed expressions are not naturally linked to the emotional state of the test subject. The instructor described beforehand and, perhaps, showed the actors the required expression, and then the actors repeated it in front of the camera. Even though there are certain differences in producing facial expressions on an involuntary or voluntary basis (Surakka & Hietanen, 1998), this fact did not affect the given research because the main aim of the dissertation was to find ways to detect facial features from expressive images (any) rather than defining a difference between these two types of facial reactions.

Apart from the databases described, there are more databases of systematically varied facial expressions such as the Facial Expression Database of Man Machine Interaction Group of Delft University of Technology (Pantic *et al*., 2005), the Belfast Naturalistic Database (Douglas-Cowie *et al*., 2003), the Facial Expression Databases and Enhanced Cohn-Kanade AU-coded Facial Expression Database of Intelligent System Lab[22], the Ekman-Hager Database of Direct Facial Actions (Ekman *et al*., 1999), the Frank-Ekman Database (Frank & Ekman, 1997), the University of Illinois at Urbana-Champaign (UIUC)-Chen Database (Chen, 2000). Nearly all these databases include images taken under the same conditions as those tested in the present research work. They consist of head-and-shoulders images with simple background, constrained lighting, and head rotations. In contrast to the databases used in this work, some of them may contain images with facial hair and eye-glasses. Some of these image collections are available to the research community; some of them have been commercialized and require authorisation and payment.

The main feature that unifies all these facial expression databases is a carefully controlled systematic approach taken in their creation. The expressions from these databases were first manually coded in terms of either prototypical facial displays or FACS scores. The images for which agreement was achieved among the coders were selected to constitute the database. After that each image from the database received a label corresponding to the expression displayed. These databases were created in order to support research mainly in the computer vision domain. Therefore all the databases consist of high-quality images in which facial

[22] http:/www.ecse.rpi.edu/~cvrl/database/database.html, (last accessed 23 April, 2007).

expressions (*i.e.* wrinkles and shadows originating from protrusions of the soft tissues of the face) are visible and detectable by AAEF systems.

3.5 SUMMARY

Two main approaches to facial expression categorization were presented. The emotion-based expression categorization is closely related to emotions and incorporates the existence of prototypical facial emotions (Ekman & Friesen, 1976). Muscle-based expression categorization is made in terms of AUs reflecting visually observable changes in the facial appearance brought about by the underlying facial muscle activations. This categorization approach is not directly connected to emotions. However, some AU and AU combinations can be classified into prototypical emotional categories.

Nowadays, computer vision researchers utilize to some extent both approaches in the field of AAEF. The need for either classification is mainly driven by possible application areas that will use AAEF output data. Emotion recognition, for example, among other expressive signals usually requires the detection and interpretation of a facial expression. The interpretation of the expression can be done either at the coarse level "happy" or "bored" or, alternatively, at the detailed level that includes small facial changes and enables distinction to be made between two happy expressions. On the other hand, there is an increasing demand from the developers of face detection and recognition systems to use muscle-based classification of the analysed expression. The reason for this is that it enables the analysis of small structural changes in the face which affect the performance of these systems.

In this dissertation, three databases, namely, the Cohn-Kanade, POFA, and JAFFE databases were selected to test the methods developed for face and facial landmark localization. To the best of my knowledge, these databases represent optimally the variety of facial expressions in terms of structural differences and changes in the intensity of the expressions. The other reason for selecting these particular databases was that they are publicly available. This facilitates a comparison of the current results with other face and feature detection methods.

It should also be mentioned that there is a large number of facial expressions databases, such as static and video material, which have been created in independent laboratories around the world. These databases are not easily available for testing but may be obtained on request sent to the database creators. There are also other face databases which may occasionally include expressive faces (usually a smile and open mouth of different intensities), but no systematic way of expression classification and image labelling is supported by them.

The existing databases of labelled facial expressions already facilitate comparisons between different AAEF approaches. Researchers can compare their results with those of other studies and draw conclusions from these comparisons. Some caution has to be applied while making these comparisons as different preprocessing can be done for the images from the database and different training schemes can be applied for method development, testing, and training.

Nevertheless, there is a need to create even more detailed databases of expressive faces coded in terms of AUs and prototypical facial displays (Pantic & Rothkrantz, 2000*a*). Thus, one of the challenges for the AAEF community would be to create an expressive database consisting of all single AUs and all possible AU combinations with a wide range of intensities. The creation of the database with prototypical expressions of varying structure and intensity levels would also benefit future progress in AAEF. However, the creation of such a database is a difficult task. Hager and Ekman (1995) spent three years "[…] collecting examples of appearances associated with known muscular actions, and can report that this task is tedious and difficult." They explain that this task requires the test subjects to be the FACS professionals and perform AUs singly and in combinations "accurately, without moving unwanted muscles, repeatedly and over multiple sessions". After that the videotapes have to be scored and labelled so that they contain "only the correct actions and nothing else". The scores have to be created not only for structure of the expression, but also its intensity value and dynamics. The same difficulties apply to the specification of the training sets of facial expressions. These databases have to be large and accurate enough to represent facial behaviour under expression variations. Another challenge relates to the development of techniques for image acquisition (*e.g.* multiple cameras for frontal and profile views), protocols, and systems which would help to create these collections of expressive data and improve the quality of expression data.

4 A Framework for Face and Facial Landmark Localization

This chapter describes a framework designed for the task of face and facial landmark localization from static facial images and streaming video. In the course of the research, the framework was continuously and iteratively developed, implemented, and tested on carefully controlled databases of posed facial expressions coded in terms of prototypical facial displays and AUs.

The framework for face and facial landmark localization consisted of several parts. Figure 4.1 shows a detailed block structure diagram of the framework. The face was located using the bottom-up approach discussed in Chapter 2, in which the location of the face was found on the basis of facial landmarks located separately. For this reason, face localization and facial landmark localization modules shared the same architectural structure in the framework.

As the framework was based on the edge information, let us first consider edges as basic image descriptors. Our choice of using edges as the basic image descriptors was based on several considerations. First, it has been

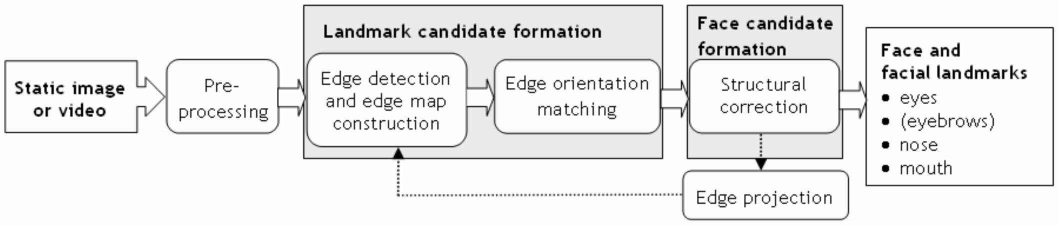


Figure 4.1. Block structure diagram of the major components of the framework for face and facial landmark localization.

shown (Bruce & Humphreys, 1994) that in the task of distinguishing between face and non-face objects, the most essential part of face information is represented by edges rather than colours, textures, and shadings. Thus the analysis of edge structures used in this dissertation had an apparent relevance to the systems of low-level pre-attentive biological vision (Marr, 1982; Rybak *et al.*, 1990; 2005). Second, from the computer vision point of view, edges represent the main discontinuities in image intensity. Edges provide a meaningful and measurable description of the face as they represent specific visual patterns which can be used to identify corresponding structures between images. Additionally, edge information tends to be robust under small changes in lighting or related camera parameters. For these reasons, edge detection has been used extensively in various computer vision tasks (Gonzalez & Woods, 2001). As will be demonstrated further, a selection of local oriented edges as basic image features was successfully applied to the task of face and facial landmark localization from images showing facial expressions. In the subsequent sections, each part of the developed framework is described separately in order of data processing. The functioning of the framework while receiving input in a form of either static images or streaming video is demonstrated.

4.1 PREPROCESSING

The first part of the framework performed a preprocessing of the input signal that can by either static image or video frame. In case of static image, after the image entered the system, it was preprocessed into the grey-scale multi-resolution representation. The $I = \{i_{ij}\}$ image was considered as a two-dimensional pixel array of the $X \times Y$ size. Each i_{ij} element of the array represented i intensity level of the (i, j) image pixel. If there was a colour image, it was first transferred into the grey-scale representation by averaging three RGB components at each pixel location:

$$i_{ij} = 0.299 \cdot R_{ij} + 0.587 \cdot G_{ij} + 0.114 \cdot B_{ij} . \quad (1)$$

This procedure allowed the method to be robust with respect to small illumination changes and skin colour variations. To eliminate noise and small details from the image, the image was further smoothed by the recursive Gaussian transformation:

$$i_{ij}^{(l)} = \sum_{p,q} a_{pq} \cdot i_{ij}^{l-1} , \quad i_{ij}^{(1)} = i_{ij} , \quad (2)$$

where a_{pq} is a coefficient of the Gaussian convolution; p and q define 5×5 size of the filter and "smoothness" of the image ($p, q = -2,\text{-}1,0,1,2$); i and j denote a current pixel location ($i = 0 \div X - 1$, $j = 0 \div Y - 1$); $X \times Y$ is image size in pixels; l defines a level of image resolution $(l = 1,2)$. The optimal values for filter size and number of resolution levels are defined experimentally.

Right image of Figure 4.2 shows a smoothed low-resolution image ($l = 2$) used to find those regions of the image which were more likely to contain facial landmarks. The original high-resolution image (*i.e.* left image of Figure 4.2, $l = 1$) was used to analyze the candidates of facial landmarks in more detail. In this way, the amount of information that was processed at a high-resolution level was significantly reduced.

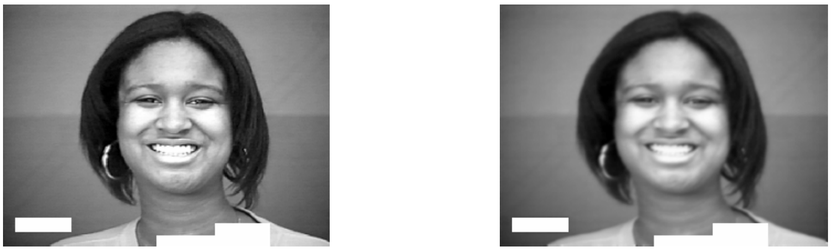


Figure 4.2. Image of happiness. Original high-resolution image (*l*=1) on the left and smoothed low-resolution image (*l*=2) on the right. The number of small details is noticeably reduced in the low-resolution image. Image labels are masked by white boxes. Courtesy of the Cohn-Kanade AU-Coded Facial Expression Database (Kanade *et al.*, 2000). Reprinted with permission.

In the case of colour video input, the face-like image region was first segmented from the background. To do this, first a procedure of histogram equalization to each video frame for all three *RGB* colour channels was applied. This calculation is not computationally expensive and is widely used to allow areas of low local contrast to gain a higher contrast without affecting a global contrast of the image (Gonzalez & Woods, 2001). After this, the original *RGB* image was transferred into YCbCr chromatic colour space as proposed in (Martinkauppi *et al.*, 2001):

$$Y_{ij} = 0.299 \cdot R_{ij} + 0.587 \cdot G_{ij} + 0.114 \cdot B_{ij}, \tag{3}$$

$$Cb_{ij} = -0.16874 \cdot R_{ij} - 0.33126 \cdot G_{ij} + 0.5 \cdot B_{ij} + 128,$$

$$Cr_{ij} = -0.5 \cdot R_{ij} - 0.41869 \cdot G_{ij} - 0.08131 \cdot B_{ij} + 128,$$

where R, G, and B are red, green , and blue components of *RGB* colour space; Y, Cb, and Cr are luminance, blue, and red components of YCbCr chromatic colour space.

Further, the skin-coloured regions of the image were subtracted from the background similarly to the method proposed in (Chang & Robles, 2000). This procedure allowed noisy regions of the image to be discarded at an early stage of processing. Therefore, it speeded up the processing and focused the following stages of the method on those parts of the image in which a face was more likely to be located. The Gaussian-fitted skin colour model was used for this purpose. The idea behind using Gaussian-fitted skin colour model was that a distribution of skin colour for different people is clustered in the YCbCr chromatic colour space and can be represented by a Gaussian distribution. The likelihood P of a skin colour for any pixel (i, j) of the image thus can be obtained with a Gaussian-fitted skin colour model.

$$P = \exp\left(-\frac{1}{2 \cdot Cv} \cdot \left((Cb - Cb_{mean})^2 + (Cr - Cr_{mean})^2 \right) \right), \tag{4}$$

where Cv is a covariance matrix; Cb_{mean} and Cr_{mean} are average red and blue chromatic components of YCbCr chromatic colour space.

After this procedure, the image obtained was a grey-scale image whose values at each pixel represented a likelihood of the given pixel belonging to the skin. Pixels with higher likelihood had darker grey-scale values. Further, the grey-scale image was thresholded and transformed into a binary image that showed skin regions and non-skin regions. The parameters for thresholding were adjusted empirically using a small face dataset. Finally, the skin-coloured region obtained was cropped from the background, converted into a grey-scale representation, and the procedure of histogram equalization was applied to it. If a skin region could not be detected at this step, the following stages of the framework were applied to the whole image converted into the histogram-equalized grey-scale representation.

4.2 Edge Detection and Edge Map Construction

The second part of the framework detected local oriented edges similar to a low-vision pre-attentive edge detection function of the visual cortex (Marr, 1982; Hubel, 1995). As has been demonstrated (Rybak *et al.*, 1990; 2005), neurons of the primary visual cortex have a remarkable property of orientation selectivity. This property provides the detection of local oriented edges and definition of their orientations. According to the concept of columnar organization (Hubel & Wiesel, 1962; 1974; Hubel, 1995), the neighbouring neurons in the primary visual cortex have similar orientation selectiveness. Together they form an orientation column or iso-orientation domain. A set of orientation columns with common receptive field forms a module of the cortex called a hypercolumn.

In the proposed framework, the edge detection module imitates hypercolumn neurons of visual cortex which are sensitive to different orientations of a local edge. The hypercolumn neurons were thought to be centred at the same point having different orientation sensitivity. The neurons interacted competitively due to strong reciprocal inhibitory interconnections in the hypercolumn. The orientation sensitivity of the "strongest neuron" encoded the edge orientation at a centre point.

To detect edges from static images, a set of multi-orientation Gaussian filters was utilized. The Gaussian filtering was selected as an edge detector because it satisfies to the criterion of "orientation selectivity", is well founded in theory, and has been successfully applied to various object detection tasks in the computer vision domain (Gonzalez & Woods, 2001). The main advantage of applying Gaussian filtering for the task of local oriented edge detection is that it gives "rich" edge structure (*i.e.* thick edges) and provides both contrast magnitude and orientation of the edge at each pixel location simultaneously (Hypermedia Image Processing Lab).

The method used to detect local oriented edges was similar to that proposed in (Golovan *et al.*, 2000). First, the smoothed low-resolution image ($l = 2$) was convolved with a set of convolution kernels which encoded different orientations and resulted from the differences of two oriented Gaussian filters with shifted centres (Equations 5 and 6). In contrast to previous studies (Golovan *et al.*, 2000; 2001; Shaposhnikov *et al.*, 2002), in this dissertation the number of edge orientations used for constructing edge maps of the image was reduced from 16 to 10. In particular, the orientations marked as $2 \div 6$ and $10 \div 14$ in Figure 4.3 were used to detect facial landmarks.

The maximum response of all ten kernels defined the contrast magnitude of the local edge at its pixel location (Equations 5-6). The orientation of the local edge was estimated with the orientation of the kernel that gave a maximum response in this pixel location (Equations 7-9).

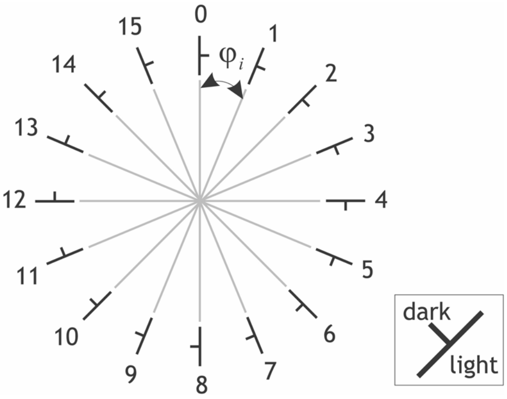


Figure 4.3. Orientation template for detection of local oriented edges, $\varphi = 22.5°$, $i = 0 \div 15$. Orientations $2 \div 6$ and $10 \div 14$ are used for orientation matching.

$$G_{\varphi_k}^{-} = \frac{1}{2 \cdot \pi \cdot \sigma^2} \cdot \exp\left(- \frac{(p - \sigma \cdot \cos\varphi_k)^2 + (q - \sigma \cdot \sin\varphi_k)^2}{2 \cdot \sigma^2} \right), \tag{5}$$

$$G_{\varphi_k}^{+} = \frac{1}{2 \cdot \pi \cdot \sigma^2} \cdot \exp\left(- \frac{(p + \sigma \cdot \cos\varphi_k)^2 + (q + \sigma \cdot \sin\varphi_k)^2}{2 \cdot \sigma^2} \right), \tag{6}$$

$$g_{ij\varphi_k} = \sum_{p,q} b^{(l)}_{i-p,j-q} \cdot G_{\varphi_k} \,, \tag{7}$$

$$G_{\varphi_k} = \frac{1}{Z} \cdot (G^-_{\varphi_k} - G^+_{\varphi_k}) \,, \tag{8}$$

$$Z = \sum (G^-_{\varphi_k} - G^+_{\varphi_k}) \,, \quad G^-_{\varphi_k} - G^+_{\varphi_k} > 0 \,. \tag{9}$$

where $\sigma = 1.2$ is a root mean square deviation of the Gaussian distribution; φ_k is the angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$, $k = 2,3,4,56,10,11,12,13,14$; p and q denote 7×7 size of the filter, $p,q = -3,\text{-}2,\text{-}1,0,1,2,3$.

Preliminary tests indicated that the use of a larger number of resolution levels or larger size of the filters did not contribute noticeably to the overall accuracy of the system. This means that the performance of the framework did not improve when values for these parameters were increased. Figure 4.4*b* demonstrates raw local oriented edges detected from the facial image. Next, the detected edges were first thresholded according to their contrast and grouped into regions of interest presumed to contain facial landmarks. The threshold for contrast filtering was determined as an average contrast of the whole smoothed low-resolution image. Edge grouping was based on the neighbourhood distance (D_n) between edge points and limited by a minimum number of edge points in the region ($N_{\min}$). Thus, edge points were grouped into one region if the distance between them was less than D_n pixels and the number of edge points inside the region was greater than $N_{\min}$. The regions with a small number of edge points were eliminated. The optimal thresholds for edge grouping were determined empirically using a small image set taken from the database. This way, the final edge map of the image consisted of regions of connected edge points representing candidates for facial landmarks.

For a more detailed description of the extracted edge regions, edge detection, contrast thresholding, and edge grouping were applied to the original high-resolution image ($l = 1$) within the limits of the located landmark candidates. In this case, the threshold for contrast filtering was determined as a double average contrast of the high-resolution image.

Figure 4.4*c* demonstrates a final edge map of the facial image. As seen from the figure, the edges were mainly concentrated around the prominent facial landmarks like eyebrows, eyes, lower nose, and mouth. There were also noisy edges in the regions of hair, decoration, face outline, wrinkles (*e.g.* nasolabial furrow caused by expression of happiness), and closing.

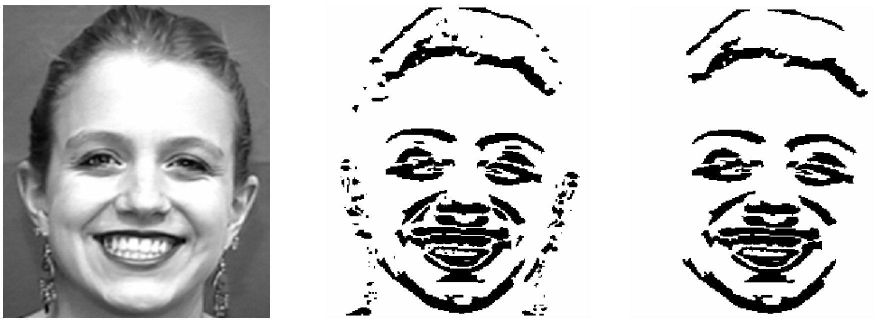


Figure 4.4. Edge detection and edge map construction: (a) original image of happiness; (b) detected local oriented edges (black regions); and (c) grouped edges after contrast thresholding. Courtesy of the Cohn-Kanade AU-Coded Facial Expression Database (Kanade *et al.*, 2000). Reprinted with permission.

If the system got a video input, the edge detection stage differed from that described above. In this case, the cropped face-like image was convolved with a 3×3 Sobel filter (Gonzalez & Woods, 2001; Hypermedia Image Processing Lab) in order to extract local oriented edges at 10 orientations. The choice of this edge detector was based on the fact that it is much faster to compute compared to Gaussian filtering. Still, it produces rich edge structures due to some "smoothing effect" of the Sobel operator. To satisfy the requirement of fast frame processing, no additional smoothing was applied at this stage. The edge points were further recursively grouped together to form regions of interest which represented candidates for facial landmarks. The shapes of the bounding boxes placed over the regions of interest located were analysed next. If the height of the bounding box was much longer than the width, this candidate was obviously not a landmark and was eliminated. Finally, the edge map of each processed video frame consisted of regions of connected edges presumed to contain facial landmarks. Some noisy edge regions persisted and needed to be eliminated. It is described below how these noisy regions were successfully discarded using the edge orientation model.

4.3 Edge Orientation Matching

After the preceding stages of edge detection and edge map construction, among the located landmark candidates there were still many noisy regions, for example, elements of hair, decoration, and clothing (as shown in Figure 4.4*c*) which had to be eliminated. In order to verify the presence of the landmark in the image or video frame, this part of the proposed framework analysed local properties of the located region of interest.

Publication I introduced the term *orientation portrait* to define a distribution of local oriented edges inside the located edge region. In order to select regions of facial landmarks, the detected edge regions were matched against the edge orientation model. The model was built on the observation that a face in the input image was usually oriented so that the eyes were at the top of the image, the mouth was at the bottom, and the nose was located in between eyes and mouth (*e.g.* video frame from a web camera). Intensive simulations demonstrated that in this case the landmark orientation portraits were constituted primarily of horizontal and nearly horizontal edges. This is visible in Figure 4.5 illustrating the averaged landmark orientation portraits for eyes regions, nose, and mouth. Therefore, orientations $2 \div 6$ and $10 \div 14$ from Figure 4.3 were used for the purpose of edge orientation matching.

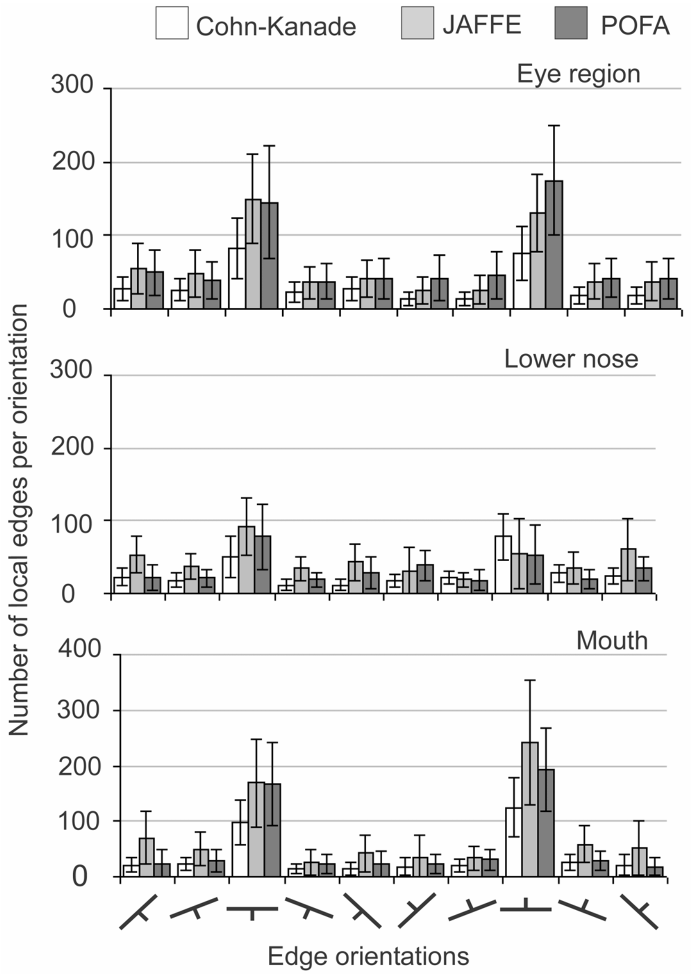


Figure 4.5. Examples of landmark orientation portraits averaged over all datasets. The error bars show plus/minus one standard deviation from the mean values. Reprinted with permission.

The characteristic property of the average landmark orientation portraits discovered was used to construct the edge orientation model. The following rules defined a structure of the *edge orientation model:* 1) horizontal orientations (*i.e.* orientations 4 and 12 from Figure 4.3) were represented by the greatest number of detected edge points; 2) the number of edges that corresponded to each horizontal orientation was more than 50% greater than the number of edges that corresponded to any other orientation; and 3) orientations could not be presented by zero number of edges.

Publication I further demonstrated that facial expressions of varying intensity and structural configuration did not affect the structure of individual landmark orientation portraits as predefined by the edge orientation model. Figure 4.6 shows orientation portraits of facial landmarks and noisy edge regions from faces with facial expressions of different structure and intensity levels. Because noisy regions typically had arbitrary distribution of the oriented edges they were discarded by the edge orientation model. As shown in Publication I, at least half of the noisy landmark candidates were successfully eliminated while applying the proposed procedure of edge orientation matching.

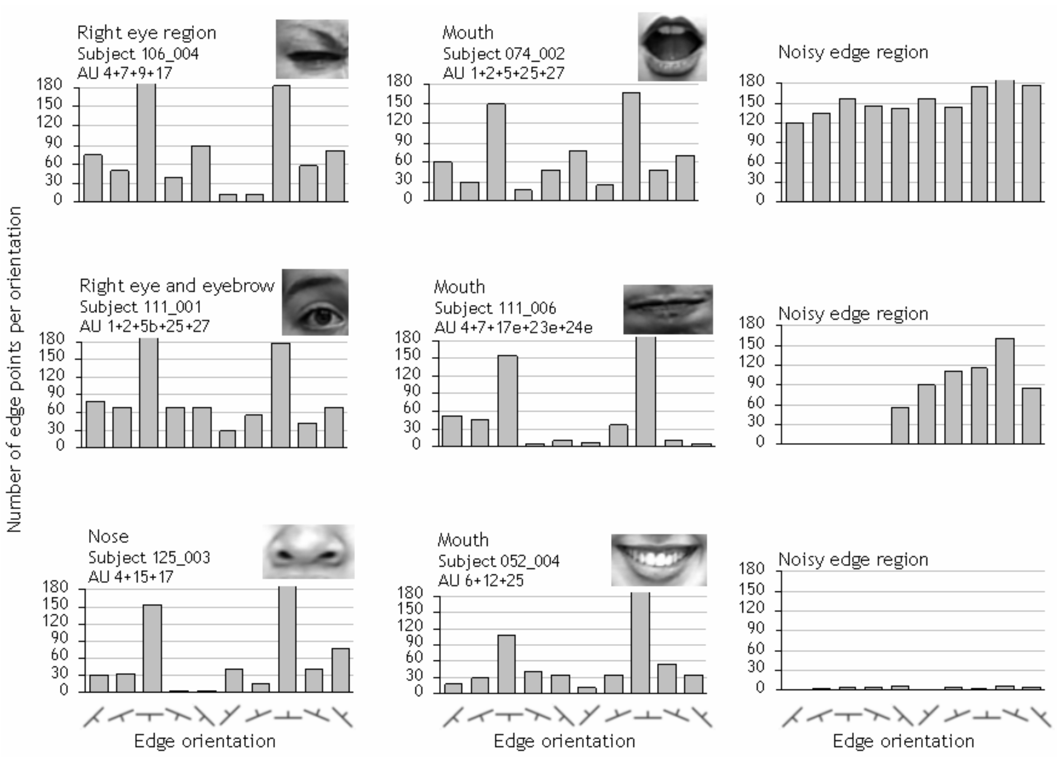


Figure 4.6. Examples of individual orientation portraits of facial landmarks from images of Cohn-Kanade facial expression database.

At the latest steps of the framework development (Publications IV-VI), the landmark candidates were allowed to have some deviations from the edge orientation model. This means that an orientation portrait of the candidate could differ slightly from the model, for example, it could have some orientations represented by zero number of edges. In further analysis,

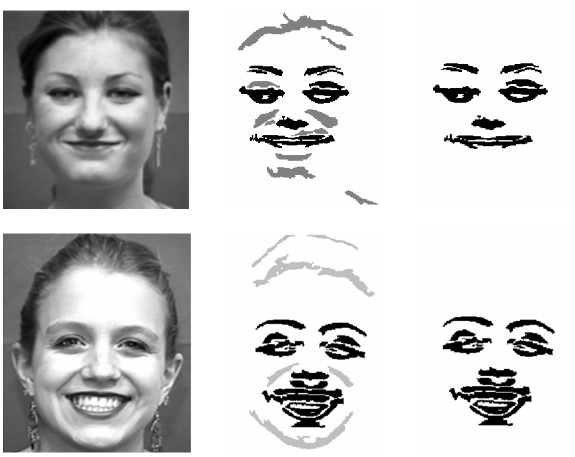


Figure 4.7. From left to right: original image of happiness; edge regions accepted (black rectangles) and discarded (grey areas) by the edge orientation model; final edge map after edge orientation matching. Courtesy of the Cohn-Kanade AU-Coded Facial Expression Database (Kanade *et al.*, 2000). Reprinted with permission.

these edge regions were also considered in composing face-like constellations of the located landmark candidates, if there were missing landmarks. Figure 4.7 shows the final edge maps of the image with landmark candidates and discarded by the edge orientation model noisy edge regions. The landmark candidates which completely corresponded to the edge orientation model were called trustable and those with deviations from the model were called supportive (i.e. support set).

4.4 STRUCTURAL CORRECTION

Structural correction was the refining part of the framework. It aimed at automatic classification of the located landmark candidates, the location of erroneous or missing landmarks. It also performed validation of the face-like constellations formed from the located landmark candidates on the basis of facial configurational information. Generally, in face and landmark localization, the analysis of spatial semantics among the located facial parts is widely used (*e.g.* Cristinacce & Cootes, 2003; 2006).

In the proposed framework, structural correction verified not only the spatial relation between separate landmark candidates, but also validated and modified landmark orientation portraits. This part of the framework is therefore called a *structural correction* rather than spatial correction. The need for development of the structural correction part became obvious as soon as it was realized that the majority of errors in landmark localization were due to erroneous merging of separate facial landmarks into one region at the early stage of edge map construction. The top row of Figure 4.8 shows examples of merged facial landmarks.

As the results of the method testing (described in Chapter 5) showed, the main reason for landmark merging was in specific changes of facial appearance caused mainly by expressions of happiness, anger, and disgust. To correct this problem, the procedure of edge projection was developed.

The schematic interpretation of the proposed technique is illustrated in Figure 4.9.

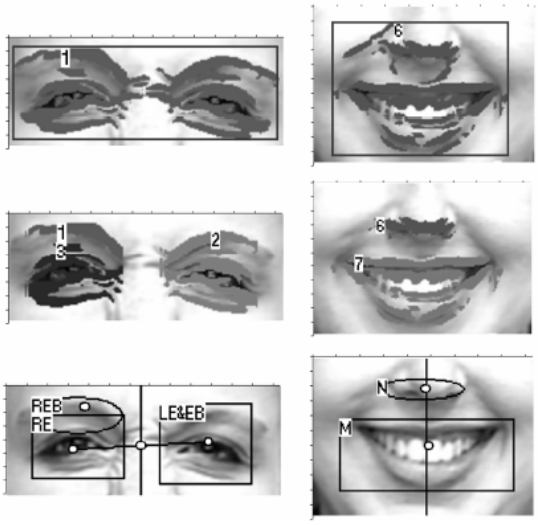


Figure 4.8. Top row shows landmarks erroneously grouped into one region. Middle row shows landmarks separated by the procedure of edge projection. Bottom row demonstrates the final detection result. Images are from the Cohn-Kanade AU-Coded Facial Expression Database (Kanade *et al.*, 2000). Reprinted with permission.

If a landmark candidate consisted of several regions of edge concentration, edge points were projected to the x-axis for upper face landmarks and to the y-axis for lower face landmarks. The projections were obtained by calculating the number of edge points along the corresponding (*i.e.* vertical or horizontal) rows of the final edge map for a given candidate. If the number of edge points was smaller than a threshold, edge points were eliminated. After each edge elimination step, if the region still was not separated the threshold was increased by 5 edge points. The initial threshold equalled to a minimum number of edges in the column (row) of a given candidate.

To classify the located landmark candidates, landmark constellations were formed and the most face-like constellations were determined. The following labels defined facial landmarks to be classified: right eye (RE),

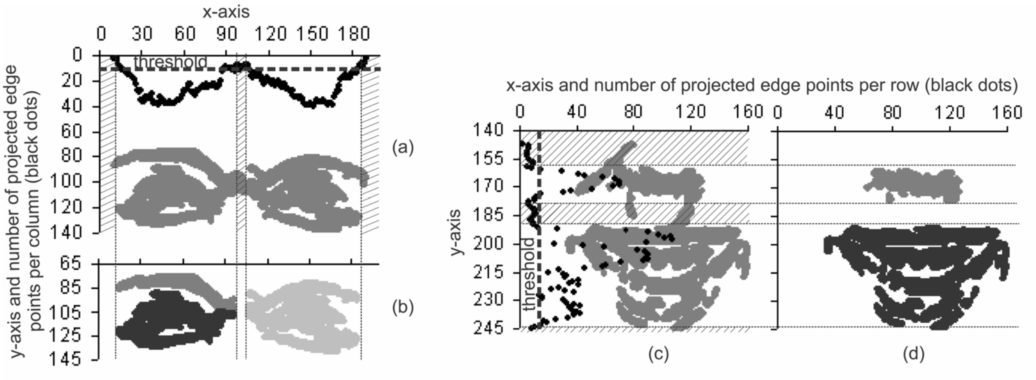


Figure 4.9. Edge maps of facial landmarks: (a) eye regions wrongly detected as one region; (b) eye regions separated by edge projection; (c) nose and mouth wrongly detected as one region; and (d) nose and mouth separated by edge projection. X- and Y-axes show the corresponding pixels in the image. Black dots represent a number of projected edge points per column or row of the upper or lower face landmark candidates respectively. Areas marked with upward diagonal lines show regions of the edge map where edges were eliminated. Reprinted with permission.

right eyebrow (REB), left eye (LE), left eyebrow (LEB), right eye and eyebrow located as one landmark (RE&EB), left eye and eyebrow located as one landmark (LE&EB), lower nose (N), and mouth (M). As the eye and eyebrow could be located separately or together, they were referred to as *eye region landmarks*. In the preliminary tests, several rules of face geometry were tried and some of these results were published in the corresponding papers (Publications III, V, and VI).

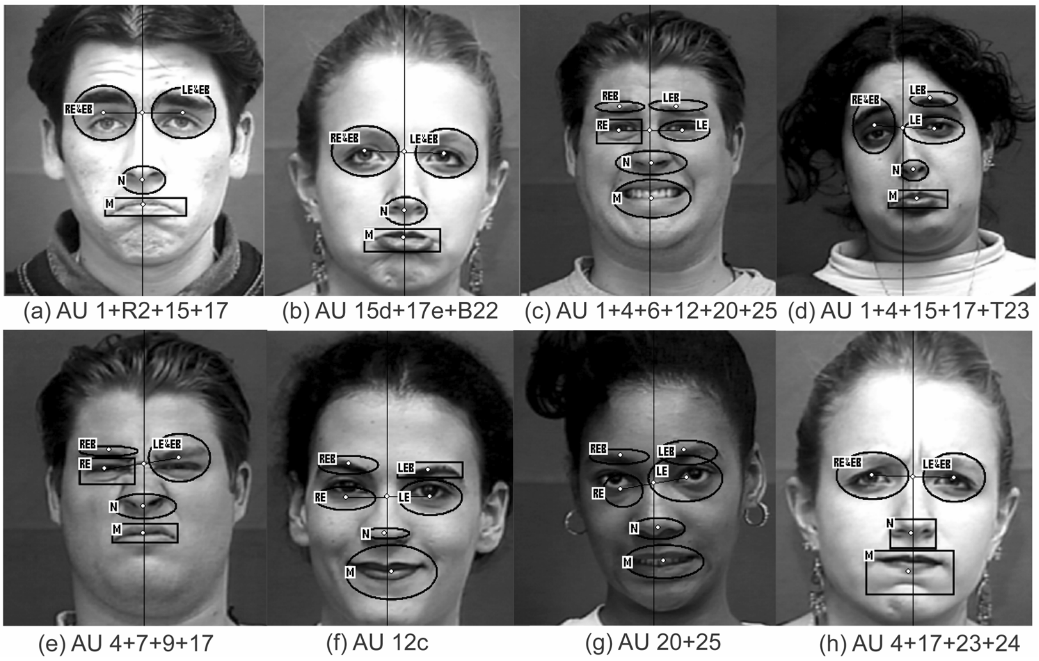
Generally, due to the side-by-side location of the upper face landmarks, they guided the entire process of the landmark classification, also those landmarks which were discarded by the orientation model. The dynamic parameter D was utilized as a measure of distances in the face model. D was calculated as the distance between mass centres of the eye region pair, so called *interocular distance.* Using this measure, the spatial constraints between locations of the rest of the candidates were verified. Other facial measures were also used for the task of landmark candidate spatial arrangement. At the same time, by utilizing geometrical relationships among the candidates, the upper face landmarks were verified. Although the method was allowed to miss landmarks, for efficient landmark detection at least one horizontal pair had to be found.

At this stage, each face candidate found consisting of four facial landmarks was given a score. The score was calculated as a sum of intermediate scores which showed how well a face candidate performed verification tests. Verification tests were fuzzy rules which defined face geometry. Each verification test was also given a weight. A number of tests was performed in order to define optimal weights for each test. This way, each test gave as its output a relative score for a given face candidate. In order to select the best-scored face-like constellation of the located landmark candidates, a new scoring function was introduced:

$$P_r = MAX - \frac{MAX}{P_{cand} / P_{\min}}, \tag{10}$$

where MAX is the maximum score for a given face candidate (we used 100); P_{cand} is a current score for a given face candidate; $P_{\min}$ is the lowest score achieved among all face candidates. This way, if we have several face-like constellations of the candidates, we will select the one that gives the highest score P_r.

In Figure 4.10 the bounding boxes placed over the facial landmarks indicate locations, and crosses indicate mass centres of the detected landmarks. The bounding boxes and crosses made up the final result of the landmark detection. The size of the bounding box depended on the size of the located landmark. The sides of the bounding box were placed at the locations of the farthest detected edges of this region in vertical and horizontal directions. After the face-like constellation of facial landmarks was found and landmark locations had been checked for correctness, the location of the face in the image was also known.


(a) AU 1+R2+15+17 (b) AU 15d+17e+B22 (c) AU 1+4+6+12+20+25 (d) AU 1+4+15+17+T23

(e) AU 4+7+9+17 (f) AU 12c (g) AU 20+25 (h) AU 4+17+23+24

Figure 4.10. Examples of the final localization results with “primary” landmarks (rectangles) and “secondary” landmarks (ovals) of the landmark localization. Images are from the Cohn-Kanade AU-Coded Facial Expression Database (Kanade *et al.*, 2000). Reprinted with permission.

4.5 Performance Evaluation

As a rule, the existing techniques of automatic face and facial feature detection evaluate the detection results as compared to human performance. Usually, the ground truth data are collected by a manual annotation of the database. The human operator marks consistently the chosen annotation features in the images or video frames. As discussed in Chapter 2, examples of the most commonly used annotation measures are the centres of the eyes, tip of the nose, and centre of the mouth.

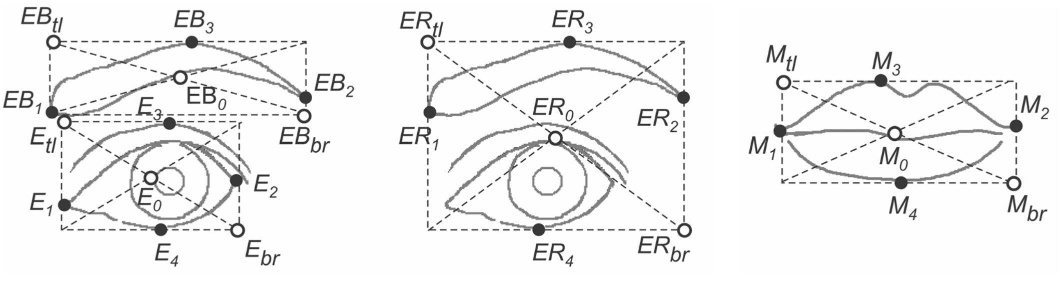
In some of the studies, in which the localization result was in the form of a point not a region, the existing performance evaluation measures were used. The point evaluation measure was adapted from (Rodriguez *et al.*, 2006). A localization result was considered correct if the distance between manually annotated and automatically detected landmark location met the following requirement:

$$d_{eye} = \frac{\max\left(d(E_R, E'_R), d(E_L, E'_L)\right)}{d(E_R, E_L)}, \tag{11}$$

where $d(a,b)$ is the Euclidean distance between point locations a and b; E_R , E_L are manually annotated and E'_R , E'_L are automatically located positions of facial landmarks. A localization was considered successful if $d_{eye} < 0.25$, which corresponds to approximately 1/4 of the annotated interocular distance $d(E_R, E_L)$ or a half of the width of the eye.

This point evaluation measure is widely used in the computer vision community and is sufficient for all applications which can make use of a single pixel point result as an output of the feature detector. However, there is a number of applications which require a feature detector to find a region of facial landmark rather than to give a point solution. To the best of my knowledge, there are no accurate criteria for the evaluation of the landmark detection result, which is represented by a region in the image instead of a single point. At the early stages of the framework development a visual inspection was utilized for this purpose. The landmark was considered correctly located if the bounding box overlapped the biggest part of the landmark and included the area surrounding the landmark less than the actual size of the landmark. In the course of the framework development, the need to develop a method of performance evaluation emerged. This would help to evaluate the results of the landmark localization fully objectively and automatically.

For this reason, a new performance evaluation measure of facial landmark detectors which output an image region as a detection result was developed. In Publication V this new measure was applied to evaluate the correctness of localization results for eye regions, nose, and mouth. In order to create a description of the landmark location in the image, the set of characteristic points shown in Figure 4.11 was selected. Four points were selected to define locations of the eye, eye region, and mouth. These points defined the right, left, top, and bottom boundaries of these landmarks. Three points were used to describe locations of the eyebrow and nose. The eyebrow location was described by its top, bottom-left, and



(a) (b) (c)

Figure 4.11. A set of characteristic points selected to define landmark location in the image: (a) eyebrow and eye; (b) eye region; and (c) mouth. The bounding box containing a landmark was defined by its top-left (*tl*) and bottom-right (*br*) bounding coordinates. The centre point of the landmark was defined by the centre point of the bounding box. Reprinted with permission.

bottom-right points. Because it was unclear how to define the vertical dimensions of the lower nose, this region was defined by the centre point of the nose tip and locations of the nostrils. All the characteristic points were manually annotated in all the databases. Further, bounding boxes were built on the base of the selected characteristic points for each landmark in all databases. The centre of the landmark was defined as the centre point of the bounding box.

The centre point, top-left, and bottom-right coordinates of the bounding box were assumed to provide a good description of the landmark location in the image and be potentially useful for the purpose of the method performance evaluation. The following pilot test was performed to verify this assumption. The characteristic points were manually annotated by 19 participants using a small dataset of images which well reflected the variations in facial expressions. The results of the pilot test supported our initial assumption; however, they also demonstrated that the characteristic points of eyebrows were difficult for humans to define and would not be as useful in the evaluation of the method performance as anticipated. In particular, some images contained subjects with very light eyebrows or eyebrows were covered by hair. For nearly all human subjects it was difficult to define boundary points for these landmarks. For some subjects, the difference in the annotation results was more than 15 pixels while the average length of the eyebrow was 38 pixels. As the aim was to compare the results of the automatic landmark localization to the human annotation results, the impact of poor definition of the eyebrow boundary points on the method performance evaluation was prevented by making the eyebrow a complementary landmark to locate. It meant that the localization performance of the upper face landmarks depended only on the localization of the eyes.

The correctness of the localization result for the nose was defined as a distance from manually annotated and automatically located centre of the nose tip. The correctness of the localization result for upper face landmarks and mouth was defined by a rectangular measure:

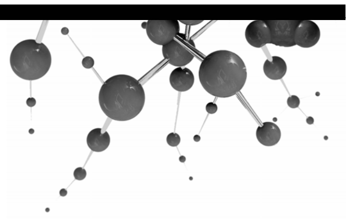
$$\max(d(p_{tl},\overline{p}_{tl}),d(p_{br},\overline{p}_{br}) \leq R\,,\ R = N \cdot StDev, \tag{12}$$

where R is a performance evaluation measure; p_{tl} , p_{br} , $\overline{p}_{tl}$, and $\overline{p}_{br}$ define the centre point, top-left and bottom-right coordinates of the manually annotated and automatically located landmark respectively; N is a number that sets a desirable accuracy for the localization result; $StDev = 2$ pixels is a standard deviation of the manual annotation averaged over all the characteristic points in the pilot test. If $\overline{p}_{tl}$ and $\overline{p}_{br}$ were found inside the manually annotated landmark position, $\overline{p}_{tl}$ and $\overline{p}_{br}$ should be located in the top-left and bottom-right quadrants of the bounding box which includes the annotated landmark.

4.6 EXECUTION TIMES

The framework can take input data in the form of both static images and streaming video. To give a rough idea of the temporal complexity of different parts of the software framework implemented (excluding visualization) in the case of static image input, the approximate running times are reported for 200×300 image size processed by the computer with a 1.7 GHz AMD Athlon TM processor and 512 MB RAM using non-optimized Visual C++ 6.0 code. Edge detection along with contrast thresholding of the detected edges at the first level of resolution took about 0.8-1 seconds per image (SPI). The computation of the final edge map at the first level of resolution together with edge grouping took about 5 SPI. The procedure of edge orientation matching took about 0.01 SPI. The running time of the edge projection part took 0.5 SPI. This is an average value, as the complexity of the edge projection computations changed depending on the number of edge projection iterations needed for landmarks to be finally separated. The structural correction of a single candidate set took 0.5 SPI on the average.

In the case of video input, the approximate running times were reported for 720×568 image size processed by the computer Dell Optiplex 745, Intel Core2 with 2.1 MHz and 1 GB DDR2 using optimised Delphi code. The software framework demonstrated high speed performance of 20-25 FPS in face localization from streaming colour video in real time.



5 Introduction to Publications

This chapter introduces six publications which represent the main scientific contribution of this dissertation. As described in the preceding chapters, the approach used combines the feature-based face and feature localization methods and a profound knowledge of how different facial muscle activations modify the appearance of a face. Accordingly, the publications consecutively describe the iterative design, implementation, and evaluation phases of the development of the face and facial landmark localization methods. Figure 5.1 illustrates the research process in a chronological order which does not correspond to the chronological order of the publications due to delays in the publication processes.

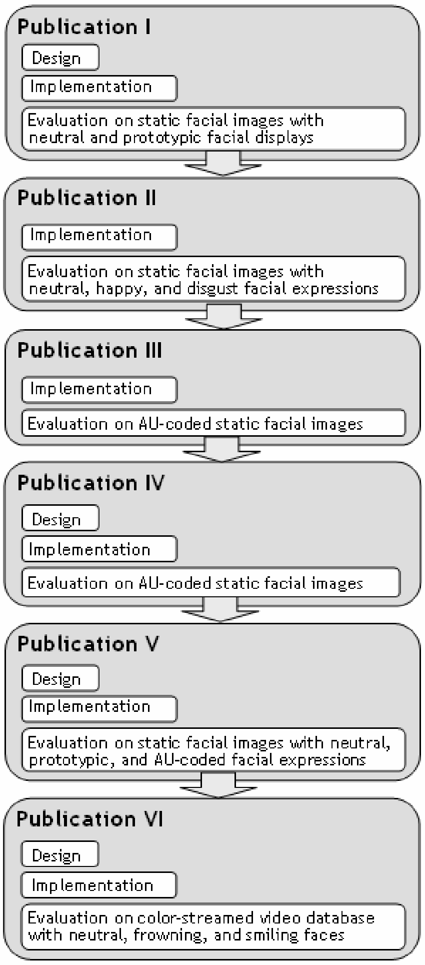

Publication I
Design
Implementation
Evaluation on static facial images with neutral and prototypic facial displays
Publication II
Implementation
Evaluation on static facial images with neutral, happy, and disgust facial expressions
Publication III
Implementation
Evaluation on AU-coded static facial images
Publication IV
Design
Implementation
Evaluation on AU-coded static facial images
Publication V
Design
Implementation
Evaluation on static facial images with neutral, prototypic, and AU-coded facial expressions
Publication VI
Design
Implementation
Evaluation on color-streamed video database with neutral, frowning, and smiling faces

Figure 5.1. The research process in chronological order denotes the design, implementation, and evaluation phases of the development of the face and facial landmark localization framework.

In the sections that follow, the topics of the individual publications are presented along with the main results from the evaluation stages of the development of the methods. Publications I-IV describe the method of facial landmark localization from static facial images. Thus, Section 5.1 describes Publication I introducing the original design of the method and presenting the results of the method performance evaluation on the database of seven prototypical facial displays. Section 5.2 describes Publication II, which aimed at finding specific facial behaviours impairing feature-based landmark localization most. In this study, the effect of lower face expressions on the method performance was investigated while testing the method on images showing neutral, happy, and disgust expressions. Looking deeper into the subject matter, Publication III described in Section 5.3, the effect of single and conjoint facial muscle activations on landmark localization was evaluated while testing the method on the AU-coded facial expression database. Section 5.4 describes how the facial behaviours discovered, which impaired landmark localization most, were taken into account while improving the overall performance of the method in Publication IV. Section 5.5 presents Publication V on the intensive evaluation of the method developed on several databases of facial expressions coded in terms of prototypical facial expressions and facial muscle activations. Finally, Section 5.6 presents Publication VI with the feature-based method of face localization from a streaming colour video.

5.1 Landmark Localization from Images Showing Prototypical Facial Expressions

Motivation and Aims of the Study: In Publication I, the original design of the feature-based method of facial landmark localization from static facial images was introduced. The aim was to design a method which is easy-to-implement and fairly robust against expressive deformations in the face.

Test Data: To test the method, the POFA database was used, which consists of 14 neutral and 96 expressive grey-scale facial images of seven prototypical facial displays: neutral, happiness, sadness, fear, anger, disgust, and surprise. The images were preset to three sizes of 100×150, 200×300, and 300×450 pixel arrays. No face alignment was performed. The potential impact of change in illumination, complex background, large head rotations, ethnicity (i.e. skin colour), occlusions, and the presence of structural components like facial hair and eye-glasses was controlled to some extent in the database. Therefore, apart from variations in facial expressions (including occlusions originating from facial expressions like bitted lips and semi-closed eyes), image size, and gender all other variables were ignored in the data analysis.

Results and Discussion: The design of the method included preprocessing, edge detection, edge map construction, and edge orientation matching steps (more details in Chapters 4.1-4.3). The first prototype was implemented to perform an experimental study and evaluate the performance of the method developed while systematically varying facial expressions and image sizes. In general, the method demonstrated a high overall performance on both neutral and expressive datasets achieving average landmark detection rates of 92% and 90% respectively. This way, the results confirmed that the choice of local oriented edges as basic image features for composing edge maps of the image was fairly robust with regard to facial expressions of varying structure and intensity. The results also showed the success of applying edge orientation matching for the elimination of false detections. After this procedure, the number of noisy landmark candidates in the image was reduced to almost a half.

More specifically, the results revealed that the chosen edge-based face representation ensured the invariance of eye localization regardless of variations in facial expression and image size. Thus, eye regions were located in 99% of the cases presented. The eye region localization was only slightly affected only by expressions of sadness and disgust. The localization of the lower face landmarks was more affected by image size and facial expressions, especially by the lower face expressions of happiness and disgust. Both, facial expressions and decrease in image size attenuated the average localization rates of mouth and nose regions to 86% and 78% respectively. The most deleterious effect on mouth and nose localization rates was caused by expressions of happiness, and disgust. Nevertheless, the within-expression variations in the mouth appearance had only a small influence on the ability of the method to find the correct mouth location. Thus, mouth was found regardless of whether it was open or closed and whether or not the teeth or tongue were visible.

The results of the first study confirmed the applicability of the method developed for the task of expression-invariant landmark localization from static frontal-view facial images. Eye region localization appeared to be robust and expression-invariant. However, nose and mouth localization needed to be improved. Some important insights into the functionality of the method developed were gained during this study. It was found that the majority of errors in nose and mouth localization occurred in the early stages of the method. The main reason for this was defined as erroneous grouping of neighbouring facial landmarks into one region. For example, expressions of happiness and disgust both caused neighbouring landmarks of the nose and mouth to merge together at the stage of edge grouping.

5.2 LANDMARK LOCALIZATION FROM IMAGES SHOWING EXPRESSIONS OF HAPPINESS AND DISGUST

Motivation and Aims of the Study: In Publication II, the aim was to continue the evaluation of the method concentrating on the lower face expressions of happiness and disgust, which impaired the landmark localization most in the previous study.

Test Data: A larger face dataset consisting of 110 images from the POFA database and 172 images from Cohn-Kanade database was chosen. The dataset included neutral, happy, and disgust expressions. This ensured a wider range of between- and within-individual variations in the structure and intensity level of the expression. The images were preset to 250×300 and 250×480 pixel arrays for the POFA and Cohn-Kanade datasets respectively. No face alignment was performed. In order to evaluate the method in a systematic and carefully controlled way, the scope of the research was narrowed down to facial expressions, having all environmental variables like image size, change in lighting, complex background, head rotations, and occlusions relatively constrained. At the same time, the effect of other variations like facial expressions, ethnicity, gender, and occlusions originated from facial expressions (*e.g.* bitted lips and closed eyes) were tested and analysed.

Results and Discussion: The same design and improved software prototype of the method from the first study were used. The results of this study corroborated the earlier findings. The procedure of edge orientation matching effectively reduced the number of noisy landmark candidates per image. The method achieved a high average localization rate of 95% for a neutral dataset. As predicted in the previous study, happiness and disgust lowered the landmark localization rates considerably. The average localization rate for the expressive dataset was decreased to 64%. In detail, although eye region localization was not affected by happiness, disgust decreased the eye region localization rate noticeably. The average localization rates for happy and disgust datasets were 100% and 67% respectively. Moreover, lower face landmarks (*i.e.* mouth and nose) were correctly located only in about a half of the happy and disgust cases. The average localization rates were 54% and 55% for nose and mouth localization respectively. The main reason for this was again the error of merging of nose and mouth landmarks into one region which occurred at the stage of edge grouping.

The combined results from Publications I and II gave an indication about facial expressions which do and do not have a deteriorating effect on the performance of the method developed. It was assumed that there would be some changes in the face which lead to the merging type of localization error. However, more detailed analysis of these facial behaviours was not possible because the facial expressions in the used databases were described in terms of prototypical facial displays. Prototypical facial displays are complex behaviours and typically include activation of multiple facial muscles. Moreover, the same expression from the database was displayed differently characterizing its within- and between-individual differences, as discussed in Chapter 3. Therefore the impact of single muscle activations on method performance was implicit, ambiguous, and excluded from the analysis. Thus, it became clear that in order to improve the design of the method, a still more detailed study was needed to accurately estimate the impact of individual muscle activations on the method performance.

5.3 Effect of Single and Conjoint AUs on Landmark Localization

Motivation and Aims of the Study: The main motivation for Publication III was the fact that deterioration in the performance of face and feature detectors due to expression variation had not been analysed in a systematic way. This way, the aim of this study was to test the method developed on AU-coded facial expression database and present the test results in a detailed and systematic way to show what specific facial muscle activations caused the degradation in the method performance.

Test Data: The method was evaluated on 468 neutral and 468 expressive images from the Cohn-Kanade database. The images were preset to 300×230 pixel arrays. No face alignment was performed. The database was classified so that it became possible to analyse the effect of single AUs and AU pairs on method performance.

Results and Discussion: The same design and somewhat improved software prototype of the method from the first study were used. The results showed that the performance of the method was considerably better on neutral than expressive datasets, which was consistent with the results from Publications I and II. The average localization rates for these two datasets were 95% and 74% respectively. The analysis of changes in the face resulting from activations of separate AUs and AU pairs clearly specified some of the critical facial behaviours which impaired the performance of the method. Expressions of happiness (*i.e.* AU 6-“cheek raiser and lid compressor” and AU 12-“lip corner puller”), disgust (*i.e.* AU 9-“nose wrinkler”, AU 10-“upper lip raiser”, and AU 11-“nasolabial furrow deepener”), anger (*i.e.* AU 4-“brow lowerer” and AU 7-“lid tightener”), and sadness (*i.e.* AU 4) caused the majority of errors.

The characteristic changes in the face were analysed when the listed AUs were displayed. For example, the lips were pulled back and obliquely upwards in the images with displayed AU 12. The displayed AUs 9 and 10 both lifted the centre of the upper lip upwards making the shape of the mouth resemble an upside down curve. AUs 9, 10, 11, and 12 all resulted in deepening of the nasolabial furrow and pulling it laterally upwards. Although there were marked differences in the shape of the nasolabial deepening and mouth shaping for these AUs, it was summed up that these AUs generally made the gap between nose and mouth smaller.

These changes in the facial appearance typically caused the error of erroneous nose and mouth merging. Similarly, AUs 4, 6, 7, and 43 and 45 (*i.e.* eye closure and blink) narrowed the space between eye lids and resulted in lowering and drawing together eyebrows and in the appearance of small wrinkles around the eye regions. All these facial behaviours caused additional contrast in the region of the nose bridge. Following this, the method detected additional noisy edges in this region. After applying the procedure of edge grouping, it usually appeared that the regions of eyes were erroneously grouped together.

The difficulty in this study was due to the fact that the database used did not present AUs alone but rather in conjunction with other supplementary AUs. Occurring singly or in combinations, these supplementary AUs typically produced skin deformations in the upper and lower face. Because of this, the effect of single AUs and AU pairs was difficult to bring into the light and only the indirect effect of AUs and their combinations on the method performance was investigated. In the course of this study the necessity of creating a systematic database consisting of all single AUs and their combinations emerged. As acknowledged in Chapter 3, the creation of such a database is a difficult and time-consuming process. However, the creation of a complete and systematic AU-coded database would greatly benefit research progress in the field of automatic face and feature detection.

On the whole, the results of this study specified some of the critical facial behaviours which caused a degradation of the method performance. The crucial facial behaviours identified gave important insights into the improvement of the method. These results were generalized to some extent to other methods of face and feature detection relying on the low-level edge and intensity information.

5.4 Landmark Localization from AU-Coded Facial Images

Motivation and Aims of the Study: The motivation for Publication IV was to improve the method in detecting lower face landmarks, particularly from images with AUs which impaired the performance of the method in the previous studies. The aim was to design and implement algorithms for the separation of the merged facial landmarks and automatic classification of the detected landmark candidates.

Test Data: The method was evaluated on 468 neutral and 468 expressive "face only" images and 468 neutral and 468 expressive "face & hair" images from the Cohn-Kanade database. The images were preset to 200×250 pixel arrays. No face alignment was performed.

Results and Discussion: The second prototype of the redesigned method allowed facial landmarks to be fully automatically located from facial images. The improvement came from the stage of structural correction (for more detail on this stage refer to Chapter 4.5) that included a procedure of edge projection to separate merged landmarks if there were any, and a proper spatial arrangement of the candidates located based on a frontal-view face geometry model.

The results of the improved method demonstrated high overall localization rates for neutral images and a wide range of expressive displays. The average localization rates for neutral and expressive datasets were 95% and 92% respectively. Compared to the results from Publication III, the introduced stage of structural correction increased the localization rates of all four facial landmarks for images with nearly all AUs and AU groups. A significant improvement in the method performance was achieved especially in the localization of lower face landmarks from images of disgust, anger, and happiness. Another important improvement of the method was that the process of facial landmark classification was now fully automatic. However, the performance evaluation of the localization results was manual and needed to be improved in order to produce a fully automatic method of face and facial landmark localization.

5.5 FULLY AUTOMATIC LANDMARK LOCALIZATION FROM EXPRESSIVE IMAGES OF HIGH COMPLEXITY

Motivation and Aims of the Study: The motivation for Publication V was to prove the applicability of the method developed to a wide range of facial expressions. The aim was to perform an extensive evaluation of the method on three facial expression databases representing facial expressions in terms of AUs and prototypical facial displays. Another aim of this study was to develop a new rectangular performance evaluation measure representing a localization result as a rectangular box instead of a conventional point representation. The main motivation behind developing a rectangular performance evaluation measure was that there were no objective criteria for the evaluation of the landmark detection result which was represented by a region in the image, instead of a single point solution.

Test Data: The method was evaluated on 468 neutral and 468 expressive images from the Cohn-Kanade database, 14 neutral and 96 expressive images from the POFA database, and 30 neutral and 176 expressive images from the JAFFE database. The images were preset to approximately 200×300 pixel arrays. No face alignment was performed. The received database represented the variability of facial expressions coded in terms of AUs and prototypical facial displays.

Results and Discussion: The third prototype of the redesigned method was evaluated by the proposed rectangular and conventional point evaluation measures (Chapter 4.5 describes the design of a new rectangular performance evaluation measure). Both types of evaluations demonstrated a high overall performance of the method. The average localization rates were 94% and 91% as evaluated by point and the new rectangular evaluation measures respectively. The results for the neutral and expressive datasets were as follows. For the neutral dataset, average localization rates were 96% evaluated by the point measure and 94% evaluated by the rectangular measure. For the expressive dataset, the average localization rates were 92% evaluated by the point measure and 88% evaluated by the rectangular measure. At this point of the method development it was realised that the performance achieved with the method evaluated by two different measures was fairly high and robust against a wide range of facial deformations represented in the publicly available facial databases. The next step in the framework development was to implement a real-time method of face and facial feature localization from a streaming video.

5.6 Real-Time Localization of Frowning and Smiling Faces Under Head Pose Variation

Motivation and Aims of the Study: The motivation of Publication VI was to expand the approach proposed in the previous studies in order to locate a face from a streaming video in real time. The aim was to adapt the algorithms developed for face and feature localization as applied to real-time image processing. The focus was fast, simple-to-compute, and expression-invariant method of face localization which can be utilized in real-time HTI applications.

Test Data: As the prospective application of the developed method was thought to lie in HTI, it was assumed that, apart from facial expressions, the input video would include head rotations. For purposes of method testing under these conditions, a systematic and controlled video database with neutral, frowning, and smiling faces under three controlled head rotations with angles of rotation of 0°, 20°, and 30° to both right and left was created. In total, the database consisted of 150 video sequences for 10 Caucasian subjects. Each sequence lasted about 8 seconds. The frame size was 720×568. No face alignment was performed.

Results and Discussion: The bottom-up face localization method was designed and implemented. The method was based on the preliminary localization of facial landmarks (described in Chapters 4.1-4.4, 4.6). After facial landmarks were located, a proper spatial arrangement of the located landmarks gave a location of the face in each video frame. The approach to facial landmark localization was similar to that proposed in Publication I.

There were also differences between these two approaches. First, the former approach worked with grey-scale images at several levels of image resolution. In contrast to this, colour-based face segmentation was proposed in this study. This allowed the face-like skin-coloured regions of the image to be extracted. These images were further transformed into the grey-scale representation. Second, a Sobel edge detector was used instead of the Gaussian edge detector. These two innovations made to the method enable real-time processing. Third, sixteen different orientations were utilized for the composition of the edge map of the image, while in the former approach the number of edge orientations was eight. The main reason for this was the difference in the nature of these two edge detectors. While the Gaussian edge detector gave rich edge structures as an output, the Sobel edge detection produced rather thin edges. In order to apply the procedure of edge orientation matching, a large number of edges were needed. This was one of the reasons for choosing a larger number of edge orientations. As in Publications IV and V, a face geometry model was used to define a proper spatial arrangement of the facial landmarks located. At the stage of face location verification, a new scoring function was introduced to select the best face-like constellation of the landmark candidates.

The method developed demonstrated sufficiently high performance in face localization from streaming colour video in real time. The speed frame of the method was about 20-24 FPS, which meets the requirement of real-time video processing defined in (Turk & Kölsch, 2004). The method was effective in the localization of faces with frontal and near-to-frontal head poses. Large head rotations decreased the localization rates. In case of frontal and near-to-frontal head postures, the method demonstrated high rates of locating faces with all three expressions tested - neutral, frowning, and smiling expressions. This gave a performance of the method for these particular facial expressions similar to the results of the previous studies (Publications IV and V).

In contrast to Publications IV and V, in this study the focus was on face localization rather than on independent landmark localization - all four facial landmarks needed to be correctly located in order to declare successful face localization. The nose was a difficult landmark to detect and produced low detection rates that resulted in low face detection rates. At the same time, some authors consider three correct landmarks sufficient to declare successful face localization (*e.g.* eyes and mouth in (Colbry *et al.*, 2005)). A choice of eyes and mouth landmarks sufficient for face detection would increase the detection rates in the method proposed. The method developed can also be applied for independent facial landmark localization if it is allowed to miss some landmarks.

6 Discussion

One of the objectives defined in the present dissertation was to find a representation of the face that remains robust regardless of a variety of changes in facial expressions. In order to address this issue, the framework proposed utilized local face representation based on local oriented edges extracted from the image or video frame. The advantage of using local face representation for purposes of efficient face and feature detection was already acknowledged in Chapter 2. Thus, it was demonstrated that local detectors outperform holistic detectors in a number of conditions such as facial expressions, head rotations, changes in lighting, and occlusions (Yow & Cipolla, 1997; Pantic & Rothkrantz, 2000b; Heisele et al., 2001; 2006; Tong et al., 2007). The choice of the edge-based approach to constructing face representation was based mainly on the fact that edge maps of the image capture the main aspects of the image (*i.e.* local discontinuities) while filtering out less important information. In addition, edge maps are well established in theory, easy to compute, compact, fairly robust under many conditions, and have been successfully applied for the task of face and feature detection in the past.

One of the key findings in the framework development was that the proposed edge-based representation of facial landmarks remained generally invariant regardless of deformations in the face brought about by facial muscle activations, including complex facial behaviours like, for example, closed eyes, bitted lips, and visible teeth. This was already evident from the results of Publication I in which an expression-invariant edge orientation model for facial landmark localization was introduced. The proposed edge orientation model defined a characteristic structure of the landmark orientation portraits (*i.e.* distribution of the local oriented edges inside the regions of facial landmarks). The results of Publication I revealed that the landmark orientation portraits generally kept the same characteristic structure as predefined by the edge orientation model for all tested facial displays. Further, Publication I corroborated the earlier findings (Golovan *et al.*, 2000; 2001; Shaposhnikov *et al.*, 2002) showing that the procedure of edge orientation matching ensured efficient differentiation between landmark and non-landmark regions in the image. This was due to the fact that the orientation portraits of noisy non-landmark image regions as a rule had an arbitrary distribution of the local oriented edges. This allowed noisy regions already to be discarded at the early stages of processing, reducing the overall computational load of the framework.

In general, Publication I served as a starting point for the whole dissertation, defining the course for further investigations. In all the subsequent studies on framework development, three main issues were addressed: 1) finding facial behaviours which impair the edge-based landmark localization, 2) taking into account the critical facial behaviours identified and searching for possible ways to improve the methods, and 3) extensive testing of the methods on various facial databases of expressive faces in order to prove the functionality of the methods developed. Publications II-VI generally confirmed that structural variability and intensity level of the expression did not affect the characteristic structure of the landmark orientation portraits. The orientation portraits of such highly deformable facial landmarks as the mouth had the same distribution of local orientation edges when the mouth was closed or opened having bitted lips or visible teeth and tongue. The region of the eyes (with or without eyebrows) demonstrated the same invariant property of having the characteristic distribution of local oriented edges under expressive deformations in the face.

In order to detect local oriented edges, a biologically plausible method was developed similar to that proposed by Golovan *et al.* (2000). As described in Chapter 4, this method imitated the pre-attentive property of "orientation selectivity" of the neurons in the primary visual cortex to give a strong response to a local edge at some particular orientation. Generally, the earlier works (Golovan *et al.*, 2000; 2001; Shaposhnikov *et al.*, 2002) which utilized a local edge-based approach to facial feature detection served as a starting point for framework development. However, in that early study no attempt was made to automatically classify the edge regions located. The issue of expression-invariant detection was neither studied nor systematically tested in these studies.

In contrast to earlier studies, in this dissertation the performance of the methods developed was systematically evaluated. This was done using several databases of facial expressions coded in terms of prototypical facial displays, like happiness and surprise, and facial muscle activations presented alone or in combinations. Three public databases of facial expressions were used, namely, the POFA, Cohn-Kanade, and JAFFE databases. Additionally, one video database was created in our laboratory and included video sequences of smiling and frowning faces under three head rotations. The databases chosen consisted of images representing a wide range of facial expressions of varying complexity. The complexity of expressions was presented by the variability of deformations in soft facial tissues (*i.e.* wrinkles and protrusions), a variety of mouth appearances including open and tight mouth, visible teeth and tongue, and self-occlusions (*i.e.* semi-closed and closed eyes and bitted lips). Apart from variations in facial expressions, the performance of the methods was tested with respect to image size variation, skin colour, out-of-plane head rotations, and the presence of clothing, decorations, and hair. However, in order to evaluate the method in a systematic and carefully controlled way the scope of the research was narrowed down to facial expressions keeping all the other affecting variables relatively constant. Thus, some limitations were placed on the input signal in the subsequent studies. For example, such destructors as change in lighting, facial occlusions by hand or hair, presence of eye-glasses or facial hair, profile head views, and cluttered scenes were generally omitted from the scope of this dissertation.

Using the databases mentioned above, the objective of the present research was to investigate the impact of facial expressions on the performance of the edge-based methods developed for face and facial landmark localization. The main motivation for doing this was the fact that only few authors (Lien *et al.*, 2000; Tian *et al.*, 2002) had reported their results referring to specific facial behaviours which impaired the detection performance. In Publication III, this issue was systematically studied by analysing the effect of single and conjoint AU activations on edge-based facial landmark localization. A number of critical facial behaviours were found which resulted in merging of neighbouring facial landmarks into one region at the early stage of edge map construction. The knowledge on the deteriorating effect of these expressive behaviours on the functioning of the localization methods developed facilitated their further improvement. Thus, a new procedure of structural correction was introduced in Publication IV, which enabled separation of the merged landmarks. The results of this study were generalized to some extent to other feature-based methods of face and feature detection relying on the low-level edge and intensity information.

As stated in Chapter 2, to date there are no objective methods of evaluation of the detection output which is represented in the form of a region in the image. It seems intuitively understandable that a point solution cannot fully describe the result, for example, of mouth localization as the mouth is a very non-rigid facial landmark. The appearance of a widely opened mouth is different from the appearance of bitted lips and it is desirable for landmark detection methods to catch this difference at the early stages of face processing. In contrast to many convenient detection methods which output a point solution (*i.e.* the centre of the landmark) defining only a position of the landmark, the localization result of the methods developed was represented by a region in the image. The dimensions and centre of the mass of the regions located enabled both defining a position of the landmark and estimating its size. In order to evaluate the accuracy of localization, the new rectangular evaluation measures were introduced in Publication V. In that study, a precise set of feature points which defined both the position and dimensions of the landmarks as rectangular boxes bounding the located landmark regions in the image was introduced. The results confirmed that the rectangular measures proposed were compact, descriptive, and robust in evaluation of the eye and mouth localization outputs in presence of facial expressions of varying structural complexity. The results of the testing of the method also showed that the methods developed revealed high rates in landmark localization as evaluated by convenient point and new rectangular measures. This way, the results demonstrated that landmark localization achieved high precision in both localizing the position of the landmark and defining its dimensions.

So far, it has been demonstrated that the proposed local edge-based face representation achieved efficient face and landmark localization. While utilizing a local face representation, the proposed framework also had elements of hybrid face representation. After the local properties of the face had been analysed, a holistic face representation was considered further to arrange the located landmark candidates according to face geometry. In Publications IV-VI, the stage of structural correction formed face-like constellations from the landmark candidates located separately and verified local similarity values for the candidates. In order to eliminate a number of false detections, the structural correction searched for facial landmarks at their expected locations in the image relative to each other (*i.e.* eyes at the top and mouth at the bottom of the image). The results from these studies showed that structural correction also enabled landmarks to be restored which had been rejected by the edge orientation model at the stage of edge orientation matching or merged with other neighbouring landmarks or hair at the stage of edge map construction.

From the start of the framework development, invariance to facial expressions was ensured in the case of eye localization. In Publications I-III, eye regions were found correctly in the majority of expressive cases regardless of whether the eyes were open, closed, or semi-closed. Because the eyes were found correctly regardless of a majority of expressive facial deformations, the locations of the eye regions found guided the search for nose and mouth landmarks in Publications IV-VI. For the present work this scheme functioned well, however, it could fail if the test conditions include, for example, an image of a complex scene. In this case, the eyes might be incorrectly detected from the very beginning. Full occlusions of the eye(s) by hand or hair would also significantly impair the performance of the method. For such conditions, it would be beneficial to use more complex models to capture the spatial arrangement of the landmark candidates detected in the image (Burl & Perona, 1996; Yow & Cipolla, 1997; Lin & Fan, 2000). The system of Salah *et al*. (2007), for example, did not assume in advance the correctness of any particular landmark location. It used subsets of all possible face-like constellations in order to validate the locations of other landmarks. The joint distribution of the landmark coordinates could be modelled with, for example, a single multi-variate Gaussian (Burl & Perona, 1996) or a mixture of bi-variate Gaussians (Salah *et al*., 2007). In this case the framework would become more robust with regard to false alarms and searching for missed or occluded landmarks at the expense of also being more computationally complex.

The framework was able to take its input either from static images or from streaming video. In the former case, when the time factor was not critical, the rather slow operation of Gaussian filtering was applied to the input image. This operation provided rich edge structures as its output meaning that a large number of edge points were detected by the Gaussian operator. It resulted in a situation in which the landmark orientation portrait consisted of a large number of points and provided a detailed description of the edge distribution inside the detected region. In the case of video input, however, the orientation portraits were built from the output of the Sobel edge detector, which provided less detailed information on the edge distribution than the Gaussian edge detector. Because some edge orientations were missing in the landmark orientation portraits in this case, it resulted in fewer local similarity measures for the detected landmark candidates. Despite this, the Sobel edge detector proved to function in real time, which is a critical requirement for many applications. Generally, there was a need to find a balance between the desired extent of details in the orientation portraits and the required speed of the framework. In the case of static input, the optimization of the code should reduce the running times significantly and extend the applicability of the algorithm for real-time video processing. The procedures of edge detection and edge grouping appeared to be the most time consuming parts of the algorithm. The reason for this was that each pixel in the image was processed by convolving it with a set of ten-orientation kernels. Altogether, this made $200 \times 300 \times 10$ single computations. To accelerate the edge detection and edge grouping parts of the framework, edge detection with a pixel grid could be applied. This would reduce the amount of pixels to be processed, therefore reducing the overall execution time of the framework. The edge grouping part can be optimized by restricting the amount of possible edge neighbours.

The landmark localization algorithms developed can be applied directly to the image if the face takes the greatest part of the image or video frame. Alternatively, the facial region can be detected first to facilitate the landmark search. If the framework takes its input from colour image data, as in Publication VI, a colour-based face pre-detection can be applied to it in order to find a region of the image in which the face is most likely to be presented. In the literature, many face and feature detection methods, for example, snake- and AAM-based methods require high-resolution images and utilize information on fine details of the face. In the framework proposed, there is no need for high-resolution images, which are not always available, especially if the input data is sourced by a low cost web camera. This is also true when the facial landmarks have to be detected from a face at a distance. There are good face detectors which can detect faces at a distance or from low resolution images (Rowley *et al.*, 1998*a*; Viola & Jones, 2001; 2004). Our method can utilize the output from such face detection systems for facial landmark localization. The limit for a size of the input facial image would be about 100x150 pixels, as reported in Publication I.

Let us now discuss the results of the present work in more detail in the context of the results from other existing face and feature detectors. In general it can be stated that the framework proposed demonstrated similar or superior performance to other face and feature detection methods in terms of localization accuracy and speed. Certainly, one needs to be cautious when comparing the performance of different face and feature detection methods. As argued in Chapter 2, it is important to take into consideration that different methods use different face databases, different preprocessing can be applied to images from the databases, and different measures can be used in order to evaluate the results of the detection output. The majority of the detection results reported in the literature were evaluated by point evaluation measures. For this reason, in this comparison the results from Publications V and VI, which have the same evaluation criteria, will mainly be used. Because the detected eye positions are sufficient criteria to declare successful face detection (Jesorsky *et al.*, 2002), the results on eye region localization from the present work will be used. It should also be mentioned that the test conditions of the face databases[23] used in some of the studies described below were different from those used in the present work. For example, some studies used the M2VTS, XM2VTS, BANCA, FRGC v.1.0, BioID, and FERET databases, which might include images with eye-glasses, facial hair, complex background, and changing lighting. These conditions were deliberately excluded in the present research as the main focus was on facial expression variation. The performance of the methods presented in

[23] The descriptions of some of the face databases mentioned in this chapter can be obtained from http://web.mit.edu/emeyers/www/face_databases.html#XM2VTSDB and http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html, (last accessed 2 September, 2008).

this work might have been different in case of unconstrained test conditions. Nevertheless, the present comparison seems to be reasonable as the majority of the databases included faces with the facial expressions tested in the present work.

First, a comparison will be made in relation to those studies which utilized local oriented edges as basic image features for face and facial feature detection and localization. A similar edge-based face detector of Fröba and Küblbeck (2002) utilized a holistic edge-based face detection scheme in which local oriented edges were applied to represent a pattern of the whole face. The reported face detection rate was 96.5% for M2VTS face database, which includes images of faces with speech articulation and simple background. The results from Publication V were consistent with these results while demonstrating the applicability of the method developed to a wider range of expressions tested.

Compared to the feature-based eye detector of Hamouz *et al.*, (2005), the present results demonstrated similar or superior rates of eye centre detection. In their study, the average detection rate of eye centre detection was more than 95% on the M2VTS database and about 79-90% on the BioID database which consists of facial images with unconstrained conditions. In the study by Song *et al.* (2006) a wavelet-based edge extraction method was proposed which combined edge and intensity information for eye centre localization. The reported performance of this method was 52.7% for the Bern database, 96.8% for AR-63 (*i.e.* AR subset consisting of neutral, smiling, and angry faces), and 86.5% for AR-564 (*i.e.* AR subset consisting of neutral, smiling, and angry faces taken under unconstrained conditions). The images from these databases had simple background and did not include faces with spectacles. The results from Publication V obtained on similar face databases demonstrated similar or higher rates for eye centre localization.

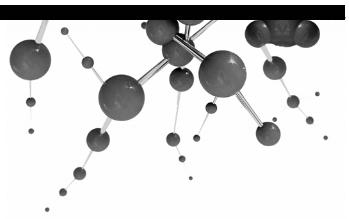
The study by Sohail and Bhattacharya (2006) utilized the generative eye detection framework originally proposed in Fasel *et al.* (2005) together with an anthropometric model of a human face. They considered eye detection in the context of successful face detection, meaning that eye locations were verified by analysing a large context of a facial region. The reported performance of the method on the JAFFE database of facial expressions was an average rate of about 99% in detecting eye centres. Their method demonstrated better precision in locating specific eye points (*i.e.* eye centres and eyelid midpoints) than the method proposed in Publication V. However, both methods were able to detect eye centres under facial expression variations resulting, for example, in closed and semi-closed eyes.

Further, the results from Publication V showed performance in eye centre localization similar to that of the existing state-of-the-art template- and learning-based face and feature detection methods. In the study by Campadelli *et al*. (2007), SVM-based feature classifiers were utilised. In their study, a two-step localization scheme was applied in which rough eye region locations were found first by classifiers trained on the database of facial images. This was followed by a more precise feature point location search by local classifier trained on the images of the eye region. The method of Cristinacce and Cootes (2006) utilized constrained local appearance models which are very similar to AAMs but involve joint shape and texture appearance models for generation of a set of region template detectors. They likewise used a two-step detection technique in order to verify local feature positions on the preliminary detected regions of facial landmarks. Thus, evaluated by point evaluation measures of 0.20-0.25 of interocular distance, the first-step rough landmark detectors of both studies were able to find on average more than 95% of eye centre points while our eye region locator was reasonably successful, finding more than 90% of eye centre points in the image. At the feature point refining step, the local feature detectors in both studies achieved an excellent performance, finding nearly 100% feature point locations inside the located landmark regions. Although, the detection rates in Cristinacce and Cootes, (2006) were given as an average for all landmarks including eyes, nose, mouth, and cheeks, the eye region localization rates can still be estimated as sufficiently high. Another learning-based feature detection method, AdaBoost-based feature classifier, by Wilson and Fernandez (2006) achieved an average eye detection rate of 93% on the Feret database, which is comparable to the performance of the eye centre localization in Publication V.

Finally, Publication VI showed that, applied to real-time video signal, the proposed framework demonstrated high speed of face localization of about 20-25 FPS. This is comparable to the speed of the existing state-of-the-art face detectors (*e.g.* Fröba & Küblbeck, 2000; Viola & Jones, 2004; Cristinacce & Cootes, 2006) and meets the requirement of real-time processing in multimodal applications (Turk & Kölsch, 2004).

The present research has several future directions. For example, there is a need to perform a comparative study and analyze the robustness of the developed and other state-of-the art face and feature detection methods in presence of complex facial expressions. The same testing conditions and face databases used would provide a solid basis for a fair comparison of the test results. Another possible research direction is to use a combination of the proposed edge-based approach and, for example, learning approach in order to detect face and its features from images of expressive faces. Further, the applicability of the proposed rectangular evaluation measures still implies their further improvement. Some important insights have been made into their development in Publication V, but further processing is still needed. For example, more extensive study should be undertaken to verify and, possibly, to improve the proposed set of feature points for defining landmark positions and dimensions in case of different facial expressions.

In summary, the results of the present dissertation demonstrated a high performance with the framework developed for the task of automatic and expression-invariant localization of the face and prominent facial landmarks such as eyes, eyebrows, nose, and mouth from static facial images and real-time video. The area of applicability of the proposed framework lies in the automatic localization of the face and facial feature locations from facial images and video sequences showing facial expressions. Emphasizing simplicity, high speed, and low computational cost of the developed algorithms, I conclude that they can be widely utilised in various applications of automatic face analysis like face identification, facial expression recognition, and vision-based HTI which typically require preprocessing of the facial information. The method can be used, for example, as a preliminary face or landmark locator that allows facial parts to be located for their subsequent analysis while eliminating irrelevant regions from the image. More specifically, there is a number of face and feature detection methods, for example, snakes, which can detect a precise contour of the face and facial landmarks from images with facial expressions. These methods require preliminary (sometimes manual) initialization of the feature locations. After the approximate locations of the features are known, these methods perform effectively. On the other hand, there are methods of face recognition and facial expression analysis which are based on the analysis of local properties of regions of facial landmarks and also require a precise initialization of the landmark locations. In these methods, the developed face and landmark locators can be utilized in preliminary localization of facial regions for their subsequent processing in which coarse localization is followed by fine feature detection.



7 Conclusions

The dissertation at hand demonstrated the general applicability of the methods developed for the task of automatic and expression-invariant localization of the face and prominent facial landmarks such as eyes, eyebrows, nose, and mouth. The six scientific publications describe the course of the iterative development, testing, and evaluation of edge-based methods of face and facial landmark localization from images containing complex facial expressions.

The methods developed combined into a framework allowing face and facial landmarks to be located automatically and efficiently from images with facial expressions of varying complexity. In addition to robustness to facial expressions, skin colour, and small in- and out-of-plane head rotations, the framework demonstrated robustness to such noise as hair and elements of clothing and decoration.

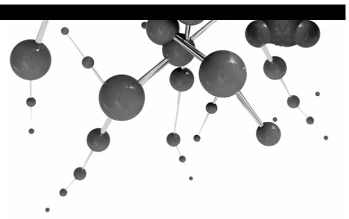
The main contributions of this dissertation are as follows.

- It was shown that the choice of the local oriented edges as basic image features for compact and descriptive face representation provided automatic face and landmark localization regardless of deformations in the face due to various facial muscle activations, including complex facial behaviours (*e.g*. closed and semi-closed eyes, open mouth with visible teeth and tongue, bitted lips, *etc.*).
- An expression-invariant edge orientation model for facial landmark localization was designed. The edge orientation model ensured expression-invariant facial landmark detection as the characteristic distribution of the local oriented edges inside the landmark regions remained the same as opposed to noisy non-landmark image regions. This property of the edge orientation model allowed the noisy regions

to be discarded in the early stages of the processing, thus significantly improving the differentiation between face and non-face objects in the image.

- New rectangular measures for the evaluation of the landmark localization outputs in the form of a rectangular box were designed and tested. The results showed that the proposed rectangular measures were compact, descriptive, and robust in the evaluation of the eye and mouth localization outputs in the presence of facial expressions of varying structural complexity.
- A number of carefully controlled empirical studies was performed in order to test the developed methods on multiple databases of expressive images and video. The results showed that the methods demonstrated high rates of face and landmark localization from images of neutral and expressive faces.
- The effect of critical facial behaviours which impaired the performance of the methods developed most, was systematically studied. This made it possible to improve the localization methods iteratively, until efficient and accurate landmark localization for these most deteriorating facial behaviours was achieved.
- Finally, a software framework for face and facial landmark localization from static images and streaming video was designed and implemented.

The proposed approach introduced a simple and compact edge-based face representation which demonstrated invariance against deformations in the facial appearance during facial expression variation. Given the robustness, simplicity, and high speed of the developed methods, it can be concluded that they are appropriate for use in automatic and expression-invariant face and facial landmark localization as such. They can also be used in the preliminary localization of facial regions for their subsequent processing in applications where coarse localization is followed by fine feature detection.



8 Bibliography

Ahonen T., Hadid A., and Pietikäinen M. (2006). Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, IEEE Computer Society, 28 (12), 2037- 2041.

Anttonen J. and Surakka V. (2005). Emotions and Heart Rate while Sitting on a Chair. *Proc. ACM Conf. Human Factors in Computing Systems (CHI'05)*, 491-499.

Augusteijn M. Skujca T. (1993). Identification Of Human Faces through Texture-Based Feature Recognition and Neural Network Technology. *Proc. IEEE Int. Conf. Neural Networks (ICNN'93)*, 1, 392-398.

Ban S.-W., Lee M., and Yang H.-S. (2004). A Face Detection Using Biologically Motivated Bottom-Up Saliency Map Model and Top-Down Perception Model. *Neurocomputing*, 56, 475-480.

Ban S.-W. and Lee M. (2005). Biologically Motivated Visual Selective Attention for Face Localization. *Lecture Notes in Comp. Science*, 3369, 196-205.

Bassili J.N. (1979). Emotion Recognition: The Role of Facial Movement and Relative Importance of Upper and Lower Areas of the Face. *J. Personality and Social Psychology*, 37, 2049-2059.

Beale J. and Keil F. (1995). Categorical Effects in the Perception of Faces. *Cognition*, 57, 217-239.

Belhumeur P.N., Hespanha J.P., and Kriegman D.J. (1997). Eigenfaces versus Fisherfaces: Recognition using Class Specific Linear Projection. *IEEE Trans. Pattern Analysis and Machine Intelligent*, 19, 711-720.

Bileschi S.M and Heisele B. (2003). Advances in Component-Based Face Detection. Pr*oc. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, 149- 156.

Black M. and Yacoob Y. (1997). Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion. *Int. J. Comp. Vision*, 25 (1), 23-48.

Bruce V. and Young A. (1986). Understanding Face Recognition. *British J. Psychology*, 77, 305–327.

Bruce V. and Humphreys, G. (1994). Recognizing Faces. *In V. Bruce and G.W. Humphreys (Eds.), Object and Face Recognition*, U.K.: Lawrence Erlbaum Associates Press, 141-180.

Burl C. and Perona P. (1996). Recognition of Planar Object Classes. *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'96)*, 223 - 230.

Calder A., Young A., Perrett D., Etcoff N., and Rowland D. (1996). Categorical Perception of Morphed Facial Expressions. *Visual Cognition*, 3, 81-117.

Campadelli P., Lanzarotti R., and Lipori G. (2007). Automatic Facial Feature Extraction for Face Recognition. *In K. Delac and M. Grgic (Eds.), Face Recognition*, Vienna: I-Tech Education and Publishing, 31-58.

Chang H. and Robles U. (2000). Face Detection, Project report, http://www-cs-students.stanford.edu/robles/ee368/main.html, (last accessed 20 May, 2008).

Chang Y., Hu Ch., and Turk M. (2004). Probabilistic Expression Analysis on Manifolds. In *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'04)*, 520-527.

Chang W.-Y., Chen Ch.-S., and Hung Y.-P. (2007). Analysing Facial Expression by Fusing Manifolds. *Lecture Notes in Comp. Science*, 4844, 621-630.

Chen L. (2000). Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Comp. Interaction. *Ph.D. Dissertation,* University of Illinois, Department of Electrical Engineering.

Chen J., Shan Sh., Yang P., Yan Sh., Chen X., Gao W. (2004). Novel Face Detection Method Based on Gabor Features. *Proc. Chinese Conf. Biometric Recognition (Sinobiometrics'04)*, 90-99.

Chen Y., Norton D., Ongur D., and Heckers S. (2008). Inefficient Face Detection in Schizophrenia. *Schizophrenia Bulletin*, 34 (2), 367-374.

Chetverikov D. and Lerch A. (1993). Multiresolution Face Detection. Proc. Workshop on Theoretical Foundations of Comp. Vision, 130-140.

Choi W.-P., Lam K.-M., Siu W.-Ch. (1999). Robust Hausdorff Distance for Human F*ace Detection* Using Genetic Algorithm. *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS'99)*, 4, 499-502.

Colbry D., Stockman G., and Anil J. (2005). Detection of Anchor Points for 3D Face Verification. *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'05)*, 3, 118-126.

Cooray S. and O'Connor N.E. (2004) Facial Feature Extraction and Principal Component Analysis for Face Detection in Color Images. *Lecture Notes in Comp. Science*, 3212, 741-749.

Cootes T.F., Taylor C.J., Cooper D.H., and Graham J. (1995). Active Shape Models—Their Training and Application. *Comp. Vision and Image Understanding*, 61 (1), 38–59.

Cootes T.F., Edwards G.J., and Taylor C. (2001). Active Appearance Models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23 (6), 681–685.

Cristinacce D. and Cootes T. (2003). Facial Feature Detection Using AdaBoost with Shape Constraints. *Proc. 14th British Machine Vision Conf. (BMV'03)*, 231-240.

Cristinacce D. and Cootes T. (2006). Feature Detection and Tracking with Constrained Local Models. *Proc. British Machine Vision Conf. (BMV'06)*, 3, 929–938.

Dailey M., Cottrell G., Padgett C., and Adolphs R. (2002). EMPATH: A Neural Network that Categorizes Facial Expressions. *J. Cognition Neuroscience*, 14 (8), 1158-1173.

Darwin C. (1872 orig.). *The Expression of the Emotions in Man and Animals*, 3rd ed., New York: Oxford University Press, 1998.

Deco G. and Obradovic D. (1996). An Information-Theoretic Approach to Neural Computing. New York: Springer-Verlag.

Dorko G. and Schmid C. (2003). Selection of Scale-Invariant Parts for Object Class Recognition. *Proc. IEEE Int. Conf. Comp. Vision (ICCV'03)*, 634–640.

Douglas-Cowie E., Cowie R., and Schroeder M. (2003). The Description of Naturally Occurring Emotional Speech. *Proc. Int. Conf. Phonetic Sciences (ICPhS'03)*, 2877-2880.

Duchenne De Boulogne G.-B. (1862 orig.). The Mechanism of Human Facial Expression. *In R.A. Cuthbertson (Ed.), Studies in Emotion and Social Interaction,* Cambridge: Cambridge University Press, 1990.

Ekman P. (1979). About Brows: Emotional and Conversational Signals. *In J. Aschoff, M. von Carnach, K. Foppa, W. Lepenies, and D. Plog (Eds.), Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, Cambridge: Cambridge University Press, 169-248.

Ekman P. (1982). *Emotion in the Human Face,* Cambridge: Cambridge University Press.

Ekman P. (1984). Expression and the Nature of Emotion. *In K. Scherer and P. Ekman (Eds.), Approaches to Emotion,* Hillsdale, N.J.: Lawrence Erlbaum, 319-344.

Ekman P. (1989). The Argument and Evidence about Universals in Facial Expressions of Emotion. *In H. Wagner and A. Manstead (Eds.), Handbook of Social Psychophysiology,* London: Lawrence Associates Press, 143-164.

Ekman P. (1999). Basic Emotions. *In T. Dalgleish and T. Power (Eds.), Handbook of Cognition and Emotion*, Sussex, U.K.: John Wiley & Sons, 45-60.

Ekman P. and Friesen W. (1976). *Pictures of Facial Affec*t, Palo Alto, California: Consulting Psychologists Press.

Ekman P. and Friesen W. (1978). *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Actio*n, Palo Alto, California: Consulting Psychologists Press.

Ekman P. and Friesen W. (1980). *Facial Action Coding System Affect Interpretation Dictionary (FACSAID)*. face-and-emotion.com/dataface/facsaid/description.jsp, (last accessed 26 June, 2008).

Ekman P., Hager J., Methvin C., and Irwin W. (1999). *Ekman-Hager Facial Action Exemplar*s. Human Interaction Laboratory, University of California, San Francisco, unpublished data.

Ekman P., Friesen W., and Hager J. (2002). *Facial Action Coding System (FACS),* Salt Lake City, UTAH: A Human Face.

Ellis H. (1981). Theoretical Aspects of Face Recognition. *In A.W. Young (Ed.), Functions of the Right Hemisphere*, London, UK: Academic Press.

Erola J. (2008). *orig.* Kasvoalueiden reaaliaikainen paikantaminen videokuvasta. *M.Sc. Thesis*, University of Tampere, Department of Comp. Sciences.

Evreinova T., Evreinov G., and Raisamo R. (2006). Video as Input: Spiral Search with the Sparse Angular Sampling. *Proc. Int. Symposium on Comp. and Information Sciences*, 542-552.

Face Detection Homepage. www.facedetection.com, (last accessed 26 June, 2008).

Farkas S. (1994). *Anthropometry of the Head and Face.* (2 ed.), Raven, New York.

Fasel I., Fortenberry B., and Movellan J.R. (2005). A Generative Framework for Real-Time Object Detection and Classification. *Computer Vision and Image Understanding*, 98, 182–210.

Feng X, Cui J, Pietikäinen M., and Hadid A. (2005). Real time facial expression recognition using local binary patterns and linear programming. *Proc. Mexican Int. Conf. Artificial Intelligence, Lecture Notes in Computer Science, 3789*, 328-336.

Féraud R., Bernier O.J., Viallet J.-E., and Collobert M. (2001). A Fast and Accurate Face Detector Based on Neural Networks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23 (1), 42-53.

Feris R., Gemmell J., Toyama K., and Krüger V. (2002). Hierarchical Wavelet Networks for Facial Feature Localization. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG'02)*, 118-123.

Frank M. and Ekman P. (1997). The Ability to Detect Deceit Generalizes across Different Types of High-Stake Lies. *Personality and Social Psycholog*y, 72, 1429–1439.

Frank C. and Nöth E. (2003). Optimizing Eigenfaces by Face Masks for Facial Expression Recognition. Proc. Int Conf. Comp. Analysis of Images and Patterns (CAIP'03), 646-654.

Freund Y. and Schapire R. E. (1995). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Proc. European Conf. Computational Learning Theor (EuroCOLT '95)*, 23–37.

Fridlund A. (1991). Evolution and Facial Action in Reflex, Social Motive, and Paralanguage. *Biological Psychology*, 32, 3-100.

Friesen W. and Ekman P. (1983). *EMFACS-7: Emotional Facial Action Coding System*. University of California at San Francisco, unpublished manuscript.

Fröba B. and Küblbeck C. (2000). Orientation Template Matching for Face Localization in Complex Visual Scenes. *Proc. IEEE Int. Conf. Image Processing (ICIP'00)*, 251-254.

Fröba B. and Küblbeck C. (2002). Robust Face Detection at Video Frame Rate Based on Edge Orientation Features. *Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'02)*, 42-347.

Gao X.W., Anishenko S., Shaposhnikov D., Podlachikova L., Batty S., and Clark J. (2007). High-Precision Detection of Facial Landmarks to Estimate Head Motions Based on Vision Models. *J. Comp. Science*, 3 (7), 528-532.

Garcia C. and Delakis M. (2002). A Neural Architecture for Fast and Robust Face Detection. *Proc. IEEE-IAPR Int. Conf. Pattern Recognition (ICPR'02)*, 2, 44-48.

Gauthier I., Anderson A.W., Tarr M.J., Skudlarski P., and Gore J.C. (1997). Levels of Categorization in Visual Recognition Studied with Functional MRI. *Current Biology*, 7, 645-651.

Gelder de B. and Rouw R. (2001). Beyond Localisation: A Dynamical Dual Route Account of Face Recognition. *Acta Psychologica*, 107, 183–207.

Gessler S., Cutting J., Frith C., and Weinman J. (1989). Schizophrenic Inability to Judge Facial Emotion: A Controlled Study. *British J. Clinical Psychology*, 28, 19–29.

Golovan A., Yoo M.-H., and Lee S.-W.L. (2000). Pre-Attentive Detection of Perceptually Important Regions in Facial Images. *Proc. Int. Conf. Pattern Recognition*, 1, 1092-1095.

Golovan A., Shevtsova N., Podladchikova L., Markin S., and Shaposhnikov D. (2001). Image Preprocessing for Identifying the Most Informative Regions in the Facial Images. *Pattern Rec. and Image Analysis: Advances in Mathematical Theory and Applications*, 11(2), 313-316.

Gonzalez R.C. and Woods R.E. (2001). *Digital Image Processing*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Govindaraju V. (1996). Locating Human Faces in Photographs. *Int. J. Comp. Vision*, 19 (2), 129-146.

Gu L., Li S.Z., and H.-J. Zhang. (2001). Learning Probabilistic Distribution Model for Multi-View Face Detection. *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'01)*, 2, 116-122.

Gu H. and Ji G. (2005). Information Extraction from Image Sequences of Real-World Facial Expressions. *Machine Vision and Applications*, 16, 105–115.

Hadid A, Pietikäinen M., and Ahonen T. (2004). A Discriminative Feature Space for Detecting and Recognizing Faces. *Proc. IEEE Conf. Comp. Vision and Pattern Recognition (CVPR'04)*, 2, 797-804.

Hager J. and Ekman P. (1995). Essential Behavioral Science of the Face and Gesture that Comp. Scientists Need to Know. *Proc. Int. Workshop on Automatic Face and Gesture Recognition (IWAFGR'95)*, Zurich, Switzerland, http://face-and-emotion.com/dataface/misctext/iwafgr.html, (last accessed 15 April, 2007).

Hamarneh G. (2000). Image Segmentation with Constrained Snakes. *Swedish Image Analysis Society Newsletter (*SSABlaskan'00), 8, 5–6.

Hamouz M., Kittler J., Kamarainen J., Paalanen P., Kälviäinen H., and Matas J. (2005). Feature-Based Affine Invariant Localization of Faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27 (9), 1490–1495.

Hannuksela J., Heikkilä J., and Pietikäinen M. (2004). A Real-Time Facial Feature Based Head Tracker. *Proc. Int. Conf. Advanced Concepts for Intelligent Vision Systems (ACIVS 2004)*, 267-272.

Haxby J., Gobbini M., Furey M., Ishai A., Schouten J., and Pietrini P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293, 2425-2430.

Heimberg C., Gur R., Erwin R., Shtasel D., and Gur R. (1992). Facial Emotion Discrimination: III. Behavioral Findings in Schizophrenia. *Psychiatry Research*, 42, 253–265.

Heisele B., Poggio T., and Pontil M. (2000). Face Detection in Still Gray images. *A.I. memo AIM-1687,* Artificial Intelligence Laboratory, MIT, 1-25.

Heisele B., Serre T., Pontil M., and Poggio T. (2001*a*). Component-Based Face Detection. *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'01)*, 1, 657-662.

Heisele B., Serre T., Pontil M., Vetter T., and Poggio T. (2001*b*). Categorization by Learning and Combining Object Parts. *In T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, (Neural Information Processing Systems: Natural and Synthetic, NIPS)*, 14 (2), 1239-1245.

Heisele B., Serre T., Mukherjee S., and Poggio T. (2003). Hierarchical Classification and Feature Reduction for Fast Face Detection with Support Vector Machines. *Pattern Recognition*, 36 (9), 2007–2017.

Heisele B., Riskov I., and Morgenstern C. (2006). Components for Object Detection and Identification*. Lecture Notes in Comp. Science*, 4170, 225-237.

Heishman R., Duric Z., and Wechsler H. (2004). Using Eye Region Biometrics to Reveal Affective and Cognitive States. *Proc. Workshop on Face Processing in Video (FPIV'04)*, 69.

Herpers R., Kattner H., Rodax H., and Sommer G. (1995). GAZE: an attentive processing strategy to detect and analyze prominent facial regions. *Proc. IEEE Int. Conf Automatic Face and Gesture Recognition (FGR'95)*, 214-220.

Herpers R., Michaelis M., Lichtenauer K.-H., and Sommer G. (1996). Edge and Keypoint Detection in Facial Regions. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'96)*, 212–217.

Hjelmas E. and Low B. (2001). Face Detection: A Survey. *Comp. Vision and Image Understanding*, 83, 235–274.

Hoch M., Fleischmann G., and Girod B. (1994). Modeling and Animation of Facial Expressions Based on B-Splines. *The Visual Comp.*, 87-95.

Hoogenboom R. and Lew M. (1996). Face Detection Using Local Maxima. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'96)*, 334-339.

Hotta K., Kurita T., and T. Mishima. (1998). Scale Invariant Face Detection Method Using Higher-Order Local Autocorrelation Features Extracted from Log-Polar Image. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'98)*, 70-75.

Hou X.W., Li S.Z., Zhang H.J., and Cheng Q.S. (2001). Direct Appearance Models. *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'01)*, 1, 828–833.

Huang J. and Wechsler H. (1999). Eye Detection Using Optimal Wavelet Packets and Radial Basis Functions (RBFs). *Int. J. Pattern Recognition and Artificial Intelligence*, 13 (7), 1009–1026.

Hubel D.H. (1995). Eye, Brain, and Vision. *Neural Sciences*, hubel.med. harvard.edu/ b14.htm, (last accessed 13 June, 2008).

Hubel D.H. and Wiesel T.N. (1962). Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. *J. Physiology*, 160 (1), 106–154.

Hubel D.H. and Wiesel T.N. (1974). Sequence Regularity and Geometry of Orientation Columns in the Monkey Striate Cortex. *J. Comp. Neurology*, 158 (3), 267–293.

Hypermedia Image Processing Lab (HIPR2), homepages.inf.ed.ac.uk/ rbf/HIPR2/hipr_top.htm, (last accessed 10 April, 2008).

Izard C. (1971). *The Face of Emotion,* New York: Appleton-Century-Crofts.

Izard C. (1979). *The Maximally Descriminative Facial Movement Coding System (MAX),* Instructional Resource Center, University of Delaware, Newark, Delaware.

Izard C., Dougherty L., and Hembree E. (1983). *A System for Identifying Affect Expressions by Holistic Judgments (AFFEX)*, Instructional Resources Center, University of Delaware, Newark, DE.

Jaimes A. and Sebe N. (2007). Multimodal Human-Comp. Interaction: A Survey. *Comp. Vision and Image Understanding*, 108 (1-2), 116-134.

Jesorsky O., Kirchberg K.J., and Frischholz R.W. (2001). Robust Face Detection using the Hausdorff Distance. *Lecture Notes in Comp. Science*, 2091, 90-95.

Ji Q., Wechsler H., Duchowski A., and Flickner M. (2005). Special Issue: Eye Detection and Tracking. *Comp. Vision and Image Understanding*, 98 (1), 1–3.

Jones M.J. and Poggio T. (1998). Multi-Dimensional Morphable Models: A Framework for Representing and Matching Object Classes. *Int. J. Comp. Vision*, 29, 107–131.

Kakumanu P., Makrogiannis S., and Bourbakis N. (2007). A Survey of Skin-Color Modeling and Detection Methods. *Pattern Recognition*, 40 (3), 1106-1122.

Kanade T., Cohn J.F., and Tian Y. (2000). Comprehensive Database for Facial Expression Analysis. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'00)*, 46-53.

Kanwisher N., McDermott J., and Chun M.M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *J. Neuroscience***,** 17 (11), 4302-4311.

Kass M., Witkin A., and Terzopoulos D. (1988). Snakes: Active Contour Models. *Int. J. Comp. Vision,* 1 (4), 321–331.

Koenen R. (2000). *MPEG-4 Project Overview*, International Organization for Standardization, ISO/IEC, JTC1/SC29/WG11, La Baule.

Kohonen T. (1977). *Associative memory: A System Theoretic Approach.*

Kotropoulos C. and Pitas I. (1997). Rule-Based Face Detection in Frontal Views. *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing (ICASSP'97)*, 4, 21–24.

Kawaguchi T. and Rizon M. (2003). Iris Detection Using Intensity and Edge Information. Pattern Recognition, 36, 549–562.

Kurita T., Hotta K., and Mishima T. (1998). Scale and Rotation Invariant Recognition Method Using Higher-Order Local Autocorrelation Features of Log-Polar Image. Proc. Asian Conf. Comp. Vision, 89-96.

Lanitis A., Taylor C.J., and Cootes T.F. (1995). An Automatic Face Identification System Using Flexible Appearance Models. *Image and Vision Computing*, 13 (5), 393-401.

Lien J., Kanade T., Cohn J., Li C. (2000). Detection, Tracking, and Classification of Action Units in Facial Expression. *J. Robotics and Autonomous Systems*, 31, 131-146.

Lienhart R. and Maydt J. (2002). An Extended Set of Haar-Like Features for Rapid Object Detection. *Proc. Int. Conf. Image Processing (ICIP'02)*, 900-903.

Lin C. and Fan K. (2000). Human Face Detection Using Geometric Triangle Relationship. *Proc. Int. Conf. Pattern Recognition (ICPR'00)*, 2, 941-944.

Low B. and Ibrahim M. (1997). A Fast and Accurate Algorithm for Facial Feature Segmentation. *Proc. Int. Conf. Image Processing (CIP'97)*, 2, 518-521.

Lu H., Zhang W., and Yang D. (2007). Eye Detection Based On Rectangle Features And Pixel-Pattern-Based Texture Features. *Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'07)*, 746-749.

Lv X.-G., Zhou J., Zhang C.-S. (2000). A Novel Algorithm for Rotated Human Face Detection. *Proc IEEE Conf. Comp. Vision and Pattern Recognition (CVPR'00)*, 1760-1765.

Lyons M.J., Akamatsu S., Kamachi M., and Gyoba J. (1998). Coding Facial Expressions with Gabor Wavelets. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'98)*, 200-205.

Ma Y., Ding X., Wang Z., and Wang N. (2004). Robust Precise Eye Location Under Probabilistic Framework. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'04)*, 339–344.

Majaranta P. and Räihä K.-J. (2007). Text Entry by Gaze: Utilizing Eye-Tracking. *In I.S. MacKenzie and K. Tanaka-Ishii (Eds.), Text Entry Systems: Mobility, Accessibility, Universality*, 175-187.

Marr D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman (Ed.), San Francisco: Freeman Publishers.

Marr D. and Hildreth E. (1980). Theory of Edge Detection. *Proc. Royal Society of London*.

Martinkauppi J., Soriano M., and Laaksonen M. (2001). Behavior of Skin Color under Varying Illumination Seen by Dierent Cameras at Dierent Color Spaces, *Machine Vision in Industrial Inspection*, 9 (4301), 102-113.

Michel P. and Kaliouby R. (2003). Real Time Facial Expression Recognition in Video Using Support Vector Machines. *Proc. Int. Conf. Multimodal Interfaces (ICMI'03)*, 258-264.

Moghaddam B. and Pentland A. (1994). Face Recognition Using View-Based and Modular Eigenspaces. *Automatic Systems for the Identification and Inspection of Humans (SPIE'94)*, 2277, 12-21.

Mohan A., Papageorgiou C., and Poggio T. (2001). Example-Based Object Detection in Images by Components. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23, 349–361.

Morimoto C.H. and Mimica M.R.M. (2005). Eye Gaze Tracking Techniques for Interactive Applications. *Comp. Vision and Image Understanding*, 98, 4–24.

Mäkinen E., and Raisamo R. (2002). Real-Time Face Detection for Kiosk Interfaces. *Proc. Asia-Pacific Conf. Comp.-Human Interaction (APCHI'02)*, 2, 528-539.

Niu Z., Shan S., Yan S., Chen X., and Gao W. (2006). 2D Cascaded AdaBoost for Eye Localization. *Proc. Int. Conf. Pattern Recognition*, 2, 1216– 1219.

Orban G.A. (1984). *Neuronal Operations in the Visual Cortex. Studies of brain functions*. Springer, Berlin.

Osuna E., Freund R., and Girosi F. (1997). Training Support Vector Machines: An Application to Face Detection. *Proc. Int. Conf. Comp. Vision and Pattern Recognition (CVPR'97)*, 130-136.

Ojala T., Pietikäinen M., and Mäenpää T. (2002). Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Bbinary Patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24, 971–987.

Pahor V. and Carrato S. (1999). A fuzzy approach to mouth corner detection. *Proc. Int. Conf. Image Processing (ICIP'99)*, 1, 667-671.

Pal N.R. and Pal S.K. (1993). A Review on Image Segmentation Techniques. *Pattern Recognition*, 26 (9), 1277-1294.

Pantic M. and Rothkrantz J. (2000*a*). Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. Pattern Analysis and Machine Intelligent*, 22 (12), 1424–1445.

Pantic M. and Rothkrantz L.J.M. (2000*b*). Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18, 881–905.

Pantic M., Valstar M., Rademaker R., and Maat L. (2005). Web-Based Database for Facial Expression Analysis. *Proc. IEEE Int. Conf. Multimedia and Expo (ICME'05)*, 317-321.

Papageorgiou P., Oren M., and Poggio T. (1998). A General Framework for Object Detection. *Proc. Int. Conf. Comp. Vision (ICCV'98)*, 555-562.

Partala T. and Surakka V. (2003). Pupil Size Variation as an Indication of Affective Processing. *Int. J. Human Comp. Studies*, 59 (1-2), 185-198.

Partala T., Surakka V., and Vanhala T. (2006). Real-Time Estimation of Emotional Experiences from Facial Expressions. *Interacting with Comp.s*, 18 (2), 208-226. (Partala *et al.*, 2006)

Perlibakas V. (2003). Automatical Detection of Face Features and Exact Face Contour. *Pattern Recognition Letters*, 24 (16), 2977–2985.

Petrushan M.V., Samarin A.I., and Shaposhnikov D.G. (2005). FOSFI: A System for Face Image Recognition. *Pattern Recognition and Image Analysis*, 15 (2), 425-427.

Picard R.W. (1997). *Affective Computing*. M.I.T. Press, Cambridge, MA.

Picard R.W. and Klein J. (2002). Comp.s that Recognise and Respond to User Emotion: Theoretical and Practical Implications. *Interacting with Comp.s*, 14 (2), 141-169.

Popovici V. and Thiran J.-P. (2001). Higher Order Autocorrelations for Pattern Classification. Proc. Int. Conf. Image Processing (ICIP'01),3, 724-727.

Reeves B. and Nass C. (1996). *The Media Equation.* Cambridge University Press.

Reisfeld D., Wolfson H., and Yeshurun Y. (1995). Context-Free Attentional Operators: The Generalized Symmetry Transform. *Int. J. Comput. Vision*, 14, 119–130.

Reisfeld D. and Yeshurun Y. (1998). Preprocessing of Face Images: Detection of Features and Pose Normalization. *Comp. Vision and Image Understanding,* 71 (3), 413-430.

Riesenhuber M. and Poggio T. (1999). Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, 2 (11), 1019–1025.

Rinn W.E. (1984). The Neuropsychology of Facial Expression: A Review of the Neurological and Psychological Mechanisms for Producing Facial Expressions. *Psychological Bulletin*, 95 (1), 52–77.

Rodriguez Y., Cardinaux F., Bengio S., and Mariéthoz J. (2006). Measuring the Performance of Face Localization Systems. *Image and Vision Computing*, 24 (8), 882-893.

Roth D., Yang M.-H., Ahuja N. (2000). A SNoW-Based Face Detector. *Advances in Neural Information Processing Systems*, 855-861.

Rowley H.A., Baluja S., and Kanade T. (1998*a*). Neural Network-Based Face Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20 (1), 23–38.

Rowley H., Baluja S., Kanade T. (1998 *b*). Rotation Invariant Neural Network-Based Face Detection. Proc. IEEE Conf. Comp. Vision and Pattern Recognition, 38-44.

Russell J. (1994). Is There Universal Recognition of Emotion from Facial Expression? *Psychological Bulletin*, 115 (1), 102-141.

Russell J.A. and Bullock M. (1986). Fuzzy Concepts and the Perception of Emotion in Facial Expressions. *Social Cognition*, 4, 309–341.

Rybak I.A., Shevtsova N.A., Podladchikova L.N., and Sandler V. M. (1990). Modeling of Neural Organization of the Visual Cortex and Some Issues of Image Processing by Neuron-Like Networks. *In A. Holden and V. Krukov (Eds.), Neural Networks: Theory and Architecture*, 117-137.

Rybak I., Gusakova V., Golovan A., Podladchikova L., and Shevtsova N. (2005). Attentionguided Recognition Based on «What»; and «Where» Representations: A Behavioral Mode. *In L. Itti, G. Rees, and J. Tsotsos (Eds.), The Encyclopedic Vol. Neurobiology of Attention*, 663-670.

Salah A., Çınar H., Akarun L., and Sankur B. (2007). Robust Facial Landmarking for Registration. *Annals of Telecommunications*, 62 (1-2), 1608-1633.

Sakai T., Nagao M., Kidode M. (1971). Processing of Multilevel Pictures by Computer - The Case of Photographs of Human Face. *Systems Computers Controls*, 2 (3), 47-54.

Sclaroff S. and Isidoro J. (2003). Active Blobs: Region-Based, Deformable Appearance Models. *Comp. Vision and Image Understanding*, 89 (2–3), 197–225.

Schneiderman H. and Kanade T. (2000). A Statistical Method for 3D Object Detection Applied to Faces and Cars. *Proc. IEEE Conf. Comp. Vision and Pattern Recognition (CVPR'00)*, 746–751.

Serre T., Wolf L., and Poggio T. (2005). Object Recognition with Features Inspired by Visual Cortex. *Proc. IEEE Conf. Comp. Vision and Pattern Recognition (CVPR'05)*, 994–1000.

Shakhnarovich G. and Moghaddam B. (2004). Face Recognition in Subspaces. *In S.Z. Li and A.K. Jain (Eds.), Handbook of Face Recognition*, 141-168.

Shaposhnikov D., Golovan A., Podladchikova L., Shevtsova N., Gao X., Gusakova V., and Gizatdinova Y. (2002). Application of the Behavioral Model of Vision for Invariant Recognition of Facial and Traffic Sign Images, (*orig.* Применение поведенческой модели зрения для инвариантного распознавания лиц и дорожных знаков. *Нейрокомпьютеры: разработка и применение),* 7-8, 21-33, (in Russian).

Shevtsova N.A., Golovan A.V., Podladchikova L.N., Gusakova V.I., Shaposhnikov D.G. and Faure A. (2007). Estimation of Motion Parameters by Retina-Like Neural Network Model. *NeuroComp.s for Image and Signal Processing*, (in press).

Sinha P. (1994). Object Recognition via Image Invariants: A Case Study, *Investigative Ophthalmology and Visual Science*, 35 (4), 1735-1740.

Sinha P., Balas B., Ostrovsky Y., and Russell R. (2006). Face Recognition by Humans: Nineteen Results All Comp. Vision Researchers Should Know About. *Proc. IEEE*, 94 (11), 948-1962.

Sirohey S. (1993). Human Face Segmentation and Identification. *Technical Report CAR-TR-695*, University of Maryland, Comp. Vision Laboratory.

Smeraldi F., Carmona O., Bigun J. (2000). Saccadic Search with Gabor Features Applied to Eye Detection and Real-Time Head Tracking. *Image and Vision Computing*, 18 (4), 323-329.

Sobottka K. and Pitas I. (1997). A Fully Automatic Approach to Facial Feature Detection and Tracking. *Lecture Notes in Comp. Science*, 1206, 77-84.

Sohail A.S. and Bhattacharya P. (2006). Detection of Facial Feature Points Using Anthropometric Face Model. *In E. Damiani, K. Yétongnon, P. Schelkens, A. Dipanda, L. Legrand, and R. Chbeir (Eds.), Signal Processing for Image Enhancement and Multimedia Processing: Multimedia Systems and Applications*, Springer US, 189-200.

Song J., Chi Z., and Liu J. (2006). A Robust Eye Detection Method Using Combined Binary Edge and Intensity Information. *Pattern Recognition*, 39 (6), 1110–1125.

Surakka V. and Vanhala T. (2008). Emotions in Human-Computer Interaction. *In A. Kappas (Ed.). Multi-Channel Communication on the Internet*. New York: Cambridge University Press, (accepted).

Surakka V. and Hietanen J. (1998). Facial and Emotional Reactions to Duchenne and non-Duchenne Smiles. *Int. J. Psychophysiology*, 29, 23-33.

Surakka V., Illi M., and Isokoski P. (2004). Gazing and Frowning as a New Technique for Human-Comp. Interaction. *ACM Trans. Applied Perception*, 1, 40-56.

Tang X., Ou Z., Su T., Sun H., and Zhao P. (2005). Robust Precise Eye Location by AdaBoost and SVM Techniques. *Proc. Int. Symposium on Neural Networks (ISNN'05)*, 93–98.

Tian, Y.-L., Kanade, T., Cohn, J., (2002). Evaluation of Gabor Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR'02)*, 229-234.

Toennies K., Behrens F., and Aurhammer M. (2002). Feasibility of Hough-Transform-Based Iris Localisation for Real-Time-Application. *Proc. Int. Conf. Pattern Recognition (ICPR'02)*, 2, 1053-1056).

Tong F., Nakayama K., Moscovitch M., Weinrib O., and Kanwisher N. (2000). Response Properties of the Human Fusiform Face Area. *Cognitive Neuropsychology*, 17 (1-3), 257-279.

Tong Y., Wang Y., Zhu Zh., and Ji Q. (2007). Robust Facial Feature Tracking Under Varying Face Pose and Facial Expression. *Pattern Recognition*, 40 (11), 3195-3208.

Tsapatsoulis N., Karpouzis K., Stamou G., Piat F., and Kollias S. (2000). A Fuzzy System for Emotion Classification Based on the MPEG-4 Facial Definition Parameter Set. *Proc. European Signal Processing Conf. (EUSIPCO'00)*, 2137-2140.

Turk M. (2005). Multimodal Human-Comp. Interaction. *In B. Kisacanin, V. Pavlovic, and T. Huang (Eds.), Real-Time Vision for Human-Comp. Interaction*, Springer, 2005.

Turk M.A. and Pentland A. (1991). Eigenfaces for Recognition. *J. Cognitive Neuroscience*, 3, 71-86.

Turk M. and Kölsch M. (2004). Perceptual Interfaces, *In G. Medioni and S.B. Kang, (Eds), Emerging Topics in Comp. Vision*, Prentice Hall.

Ullman S., Vidal-Naquet M., and Sali E. (2002). Visual Features of Intermdediate Complexity and Their Use in Classification. *Nature Neuroscience*, 5 (7), 682–687.

Viola P. and Jones M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. *Proc. IEEE Comp. Society Conf. Comp. Vision and Pattern Recognition (CVPR'01)*, 1, 511-518.

Viola P. and Jones M. (2004). Robust Real-Time Face Detection. *Int. J. Comp. Vision*, 57 (2), 137–154.

Vukadinovic D. and Pantic M. (2005). Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. Proc. *IEEE Int. Conf. Systems, Man And Cybernetics (SMC'05)*, 1692–1698.

Walther D., Itti L., Riesenhuber M., Poggio T., and Koch C. (2002). Attentional Selection for Object Recognition—A Gentle Way. *Lecture Notes in Comp. Science*, 2525, 472–479.

Wang Y., Chua C., and Ho Y. (2002). Facial Feature Detection and Face Recognition from 2D and 3D Images. *Pattern Recognition Letters*, 23 (10), 1191-1202.

Wang P., Green M., Ji Q., and Wayman J. (2005). Automatic Eye Detection and Its Validation. *Proc. IEEE Conf. Comp. Vision and Pattern Recognition (CVPR'05)*, 3, 164.

Weber M., Welling W., and Perona P. (2000). Towards Automatic Discovery of Object Categories. *Proc. IEEE Conf. Comp. Vision and Pattern Recognition (CVPR'00)*, 2, 101-108.

Wilson Ph. I. and Fernandez J. (2006). Facial Feature Detection Using Haar Classifiers. *J. Computing Sciences in Colleges*, 21 (4), 127-133.

Wiskott L., Fellous J.M., Krüger N., and Malsburg der C.V. (1997). Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19 (7), 775–779.

Yacoob Y., Lam H-M., and Davis L. (1995). Recognizing Faces Showing Expressions. *Proc. Int. Workshop on Automatic Face and Gesture Recognition (FGR'95)*, 278-283.

Yang G.Z. and Huang T.S. (1994). Human Face Detection in a Complex Background. *Pattern Recognition*, 27 (1), 53–63.

Yang M., Kriegman D., and Ahuaja N. (2002). Detecting Face in Images: A Survey. *IEEE Trans. Pattern Analysis and Image Understanding*, 24, 34-58.

Yow K.C. and Cipolla R. (1997). Feature-Based Human Face Detection. *Image and Vision Computing,* 15 (9), 713-735.

Yuille A., Hallinan P., and Cohen D. (1992). Feature Extraction from Faces Using Deformable Templates*. Int. J. Comp. Vision*, 8 (2), 99–111.

Zhang Z., Lyons M., Schuster M., Akamatsu S. (1998). Comparison between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition using Multi-Layer Perceptron. *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (AFGR'1998)*, 454-459.

Zhang W., Shan S., Gao W., Chen X., and Zhang H. (2005). Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition. *Proc. IEEE Int. Conf. Comp. Vision (ICCV'05)*, 786-791.

Zhao W., Chellappa R., Phillips P., and Rosenfeld A. (2003). Face Recognition: A Literature Survey. *ACM Computing Surveys*, 35 (4), 399–458.

Zhao G. and Pietikäinen M. (2007). Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *Trans. Pattern Analysis and Machine Intelligence*, 29 (6), 915-928.

Zion-Golumbic E. and Bentin Shl. (2007). Dissociated Neural Mechanisms for Face Detection and Configural Encoding: Evidences from N170 and Induced Gamma-Band Oscillation Effects. *Cerebral Cortex*, 17 (8), 1741-1749.

Publication I

Gizatdinova Y. and Surakka V. (2006). Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, 28 (1), 135-139.

Available online at:
http://ieeexplore.ieee.org (requires subscription)

Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images

Yulia Gizatdinova and Veikko Surakka

Abstract—Feature-based method for detecting landmarks from facial images was designed. The method was based on extracting oriented edges and constructing edge maps at two resolution levels. Edge regions with characteristic edge pattern formed landmark candidates. The method ensured invariance to expressions while detecting eyes. Nose and mouth detection was deteriorated by happiness and disgust.

Index Terms—Computing methodologies, image processing and computer vision, segmentation, edge and feature detection.

1 Introduction

Automated detection and segmentation of a face have been active research topics for the last few decades. The motivation behind developing systems of face detection and segmentation is a great number of its applications. For example, detection of a face and its features is an essential requirement for face and facial expression recognition [1], [2], [3], [4].

Due to such factors as illumination, head pose, expression, and scale the facial features vary greatly in their appearance. Yacoob et al. [5] demonstrated that facial expressions are particularly important factors affecting automated detection of facial features. They aimed to compare the recognition performance of template and feature-based approaches to face recognition. Both approaches resulted in worse recognition performance for expressive images than for neutral ones.

Facial expressions as emotionally or otherwise socially meaningful communicative signals have been intensively studied in psychological literature. Ekman and Friesen [6] developed the Facial Action Coding System (FACS) for coding all visually observable changes in the human face. According to FACS, a muscular activity producing changes in facial appearance is coded in the terms of action units (AU). Specific combinations of AUs represent prototypic facial displays: neutral, happiness, sadness, fear, anger, surprise, and disgust [7]. At present, there is good empirical evidence and good theoretical background for analyzing how different facial muscle activations modify the appearance of a face during emotional and social reactions [8], [9], [10].

Studies addressing the problem of automated and expression-invariant detection of facial features have been recently published. In particular, to optimize feature detection some attempts have been made to utilize both profound knowledge on human face and its behavior and modern imaging techniques. Comprehensive literature overviews on different approaches to face and facial feature detection have been published by Hjelmas and Low [11] and Yang et al. [12].

Liu et al. [13] investigated facial asymmetry under expression variation. The analysis of facial asymmetry revealed individual differences that were relatively unaffected by changes in facial expressions. Combining asymmetry information and conventional template-based methods of face identification, they achieved a high rate of error reduction for face classification.

Tian et al. [14] developed a method for recognizing several specifically chosen AUs and their combinations. They analyzed both stable facial features as landmarks and temporal facial features like wrinkles and furrows. The reported recognition rates were high for recognizing AUs from both upper and lower part of a face.

Golovan [15] proposed a feature-based method for detecting facial landmarks as concentrations of the points of interest. The method demonstrated high detection rate and invariance to changes in image view and size while detecting facial landmarks. However, the method was not tested with databases of carefully controlled facial expressions. We extended the method introduced by Golovan to detect facial landmarks from expressive facial images. In this framework, the aim of the present study was to experimentally evaluate the sensitivity of the developed method while systematically varying facial expression and image size.

2 Database

The Pictures of Facial Affect database [16] was used to test the method developed for detection of facial landmarks. The database consists of 110 images of 14 individuals (i.e., six males and eight females) representing neutral and six prototypical facial expressions of emotions: happiness, sadness, fear, anger, surprise, and disgust [7]. On average, there were about 16 pictures per expression. In order to test the effects of image resizing on the operation of the developed method, the images were manually normalized to three preset sizes (i.e., 100×150, 200×300, and 300×450 pixels). In sum, $110 \times 3 = 330$ images were used to test the method.

3 Facial Landmark Detection

The regions of eyebrow-eyes, lower nose, and mouth were selected as facial landmarks to be detected. There were two enhancements to the method proposed in previous works [15], [17]. The first enhancement is the reduction of the number of edge orientations used for constructing edge maps of the image. In particular, the orientations ranging from 45 degrees to 135 degrees and 225 degrees to 315 degrees in step of 22.5 degrees were used to detect facial landmarks (Fig. 1). The chosen representation of edge orientations described facial landmarks relatively well and reduced a computational load of the method. The second enhancement is the construction of the orientation model of facial landmarks. The landmark model was used to verify the existence of a landmark in the image.

The method was implemented through three stages: preprocessing, edge map constructing, and orientation matching. These stages will be described in details in the following sections.

3.1 Preprocessing

First, an image is transformed into the gray-level representation. To eliminate noise edges and remove small details, the gray-level image is then smoothed by the recursive Gaussian transformation. The smoothed images are used to detect all possible candidates for facial landmarks, and no smoothed images—to analyze the landmark candidates in details (Figs. 2a and 2b). In that way, the amount of information that is processed at a high resolution level is significantly reduced.

3.2 Edge Map Constructing

Local oriented edges are extracted by convolving a smoothed image with a set of 10 kernels. Each kernel is sensitive to one of 10 chosen orientations. The whole set of 10 kernels results from differences between two oriented Gaussians with shifted kernels.

$$G_{\varphi_k} = \frac{1}{Z}\left(G^{-}_{\varphi_k} - G^{+}_{\varphi_k}\right), \tag{1}$$

- *Y. Gizatdinova is with the Research Group for Emotions, Sociality, and Computing, Tampere Unit for Computer-Human Interaction, Department of Computer Sciences, University of Tampere, Tampere, FIN-33014, Finland. E-mail: ig74400@cs.uta.fi.*
- *V. Surakka is with the Research Group for Emotions, Sociality, and Computing, Tampere Unit for Computer-Human Interaction, Department of Computer Sciences, University of Tampere, Tampere, FIN-33014, Finland, and he is also with the Department of Clinical Neurophysiology, Tampere University Hospital, Tampere, FIN-33521, Finland. E-mail: Veikko.Surakka@uta.fi.*

Manuscript received 2 July 2004; revised 20 Apr. 2005; accepted 5 May 2005; published online 11 Nov. 2005.
Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0333-0704.

0162-8828/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

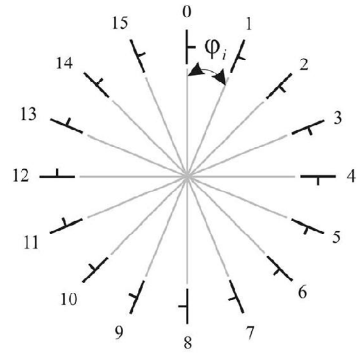


Fig. 1. Orientation template for extracting local oriented edges, $\varphi_i = i \cdot 22.5$, $i = 0 \div 15$. Edge orientations used for detecting facial landmarks were marked as numbers $2 \div 6$ and $10 \div 14$.

$$Z = \sum_{p,q} \left(G^-_{\varphi_k} - G^+_{\varphi_k}\right), G^-_{\varphi_k} - G^+_{\varphi_k} > 0, \tag{2}$$

$$G^-_{\varphi_k} = \frac{1}{2\pi\sigma^2}\exp\left(\frac{(p - \sigma\cos\varphi_k)^2 + (q - \sigma\sin\varphi_k)^2}{2\sigma^2}\right), \tag{3}$$

$$G^+_{\varphi_k} = \frac{1}{2\pi\sigma^2}\exp\left(\frac{(p + \sigma\cos\varphi_k)^2 + (q + \sigma\sin\varphi_k)^2}{2\sigma^2}\right), \tag{4}$$

where σ is a root mean square deviation of the Gaussian distribution, φ_k is angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5$, $k = 2, 3, 4, 5, 6, 10, 11, 12, 13, 14$, $p, q = -3, -2, -1, 0, 1, 2, 3$.

The maximum response of all 10 kernels defines the contrast magnitude of a local edge at its pixel location. The orientation of a local edge is estimated with the orientation of a kernel that gave the maximum response.

$$g_{ij\varphi_k} = \sum_{p,q} b^{(l)}_{i-p,j-q} G_{\varphi_k}, \tag{5}$$

where b denotes the gray level of the image at pixel $(i, j); i = 0 \div W - 1; j = 0 \div H - 1; W, H$ are, respectively, the width and height of the image, $l = 1, 2$.

The threshold for contrast filtering of the extracted edges is determined as an average contrast of the whole smoothed image. Edge grouping is based on the neighborhood distances between oriented edges and is limited by a possible number of neighbors for each edge. The optimal thresholds for edge grouping are determined using small image set taken from the database. In such a way, the edge map of the smoothed image (i.e., $l = 2$) consists of the regions of edge concentrations presumed to contain facial landmarks. Fig. 2c presents the primary feature map that was constructed by detecting local edges of 10 chosen orientations. Fig. 2d shows the primary map after contrast thresholding and grouping extracted edges into the candidates for facial landmarks.

To get a more detailed description of the extracted edge regions, edge extracting and edge grouping are applied to high resolution image (i.e., $l = 1$) within the limits of these regions. In this case, the threshold for contrast filtering is determined as a double average contrast of the high resolution image.

3.3 Orientation Matching

We analyzed orientation portraits of edge regions extracted from 12 expressive faces of the same person. On the one hand, expressions do not affect specific distribution of the oriented edges contained in regions of facial landmarks (Fig. 3a). On the other hand, noise regions have arbitrary distribution of the oriented edges (Fig. 3b).

Finally, we created four average orientation portraits for each facial landmark. Average orientation portraits keep the same specific pattern of the oriented edges as individual ones (Fig. 4).

3.3.1 Orientation Model

Such findings allowed us to design the characteristic orientation model for all four facial landmarks. The following rules define the structure of the orientation model: 1) Horizontal orientations are represented by the greatest number of extracted edges, 2) a number of edges corresponding to each of horizontal orientations is more than 50 percent greater than a number of edges corresponding to other orientations taken separately, and 3) orientations cannot be presented by zero number of edges.

The detected candidates for facial landmarks are manually classified into one of the following groups: noise or facial landmark like eye, nose, and mouth. Fig. 2e reveals the final feature map consisting of candidates whose orientation portraits match with the orientation model.

4 RESULTS

Fig. 5 illustrates examples of the landmark detection from neutral and expressive facial images.

On the stage of edge map constructing, an average number of candidates per image was 8.35 and did not vary significantly by changes in facial expression and image size. After the orientation matching, the average number of candidates per image was reduced to almost a half and amounted to 4.52. Fig. 6 illustrates the decrease

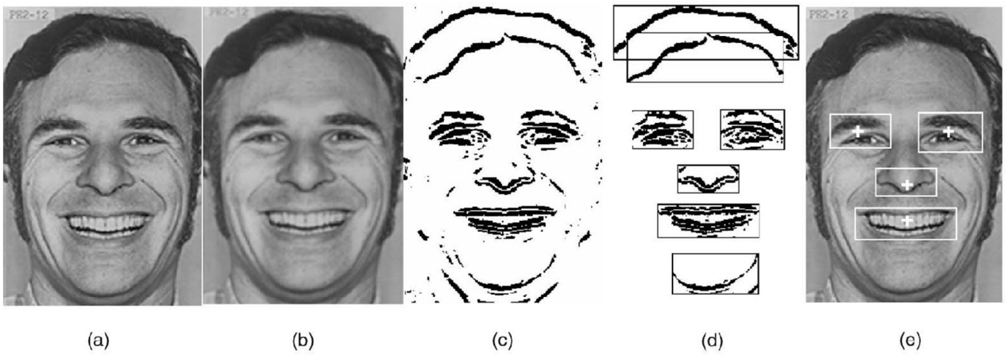


Fig. 2. Landmark detection: (a) Image of happiness, (b) smoothed image ($\sigma = 0.8$), (c) extracted oriented edges ($\sigma = 1.2$), (d) landmark candidates, and (e) facial landmarks and their centers of mass. Image from Pictures of Facial Affect. Copyright © 1976 by Paul Ekman. Reprinted with permission of Paul Ekman.

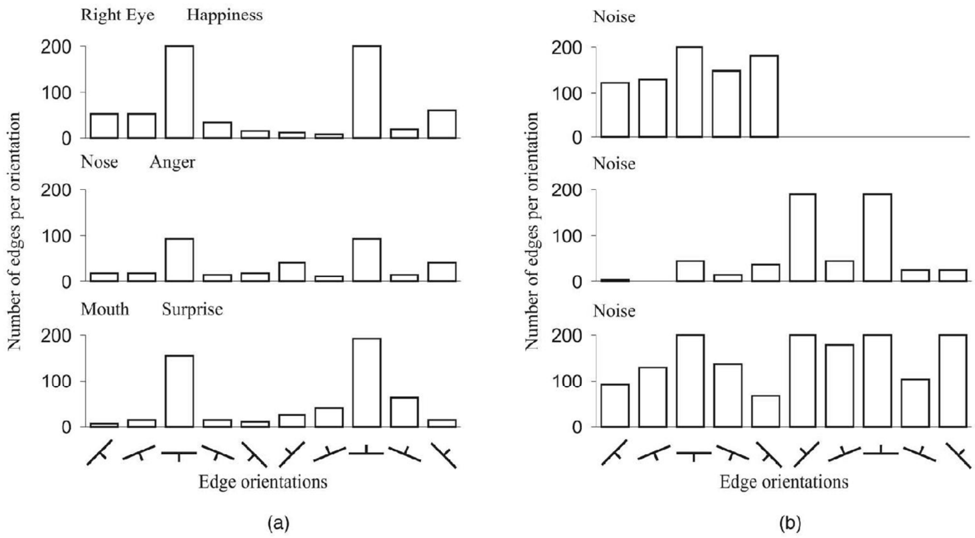


Fig. 3. Individual orientation portraits of (a) facial landmarks with specific distribution of the oriented edges and (b) noise regions with arbitrary distribution of the oriented edges.

in the number of candidates per image averaged over different facial expressions.

Table 1 shows that the developed method revealed an average detection rate of 90 percent in detecting all four facial landmarks from both neutral and expressive images. The average detection rates were 94 percent and 90 percent for neutral and expressive images, respectively. The detection of nose and mouth was more affected by facial expressions than the detection of eyes.

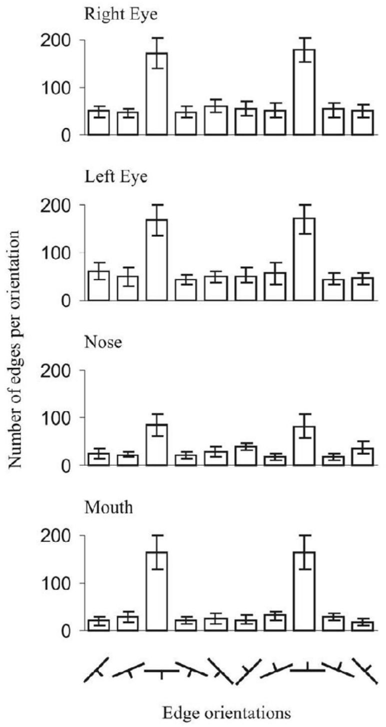


Fig 4. Average orientation portraits of landmarks with specific distribution of the oriented edges. The error bars show plus/minus one standard deviation from the mean values.

Both eyes were detected with a high detection rate from nearly all types of the images. In general, the correct detection of eyes did not require a strong contrast between the whites of eyes and iris. In such a way, eyes were found correctly regardless of whether the whites of eyes were visible or not (Fig. 5). However, expressions of sadness and disgust reduced the average detection rate to 96 percent. The correct eye localization was only slightly affected by variations in image size.

Regions of both eyes had nearly the same number of the extracted oriented edges. About one third of the total number of edges were extracted from the regions of eyebrows. As a result, the mass centers of the eye regions slightly shifted up from the iris centers (Fig. 5).

Detection of the mouth region was more affected by changes in facial expression and image size than detection of the eye regions. On average, the correct location of the mouth region was found in more than 90 percent of the expressive images with the exception of happiness (82 percent) and disgust images (49 percent). The smallest image size had a marked deteriorating effect on the mouth detection. However, the within-expression variations in the shape of the mouth had only a small influence on the ability of the method to mark the correct area. As a rule, the mouth region was found regardless of whether the mouth was open or closed and whether the teeth were visible or not.

The nose detection was even more affected by variations in facial expression and image size than the mouth detection. The expressions of happiness, surprise, and disgust had the biggest deteriorating effect on the detection of the nose region. The average detection rate for nose region was 74 percent for happiness, 78 percent for surprise, and 51 percent for disgust images. It was more than 81 percent for other expressive images. In sum, the images expressing disgust was considered the hardest to process (Fig. 5).

There were three types of errors in detecting facial landmarks. Fig. 7 gives examples of such errors. The undetected facial landmarks were considered to be the errors of the first type. Such errors occurred when a region of interest including a facial landmark was rejected as a noise region. In particular, the nose was the most undetectable facial landmark (Fig. 7a). The incorrectly grouped landmarks were regarded as the errors of the second type. The most common error of the second type was grouping regions of nose and mouth in one region (Fig. 7b and Fig. 7c). There were

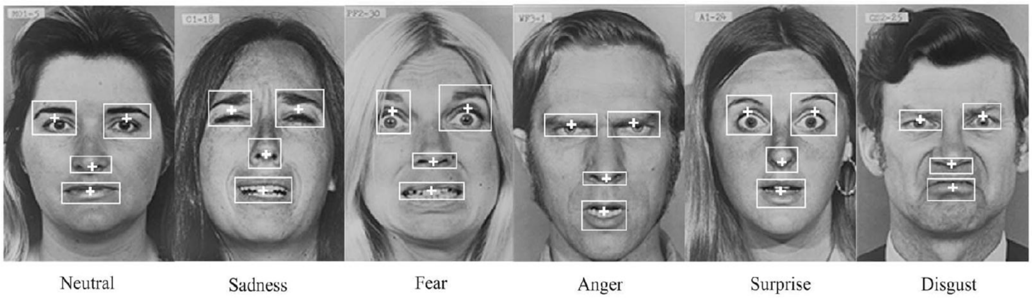


Fig. 5. Detected facial landmarks and their centers of mass. Images from Pictures of Facial Affect. Copyright©1976 by Paul Ekman. Reprinted with permission of Paul Ekman.

only a few cases of grouping together eye regions (Fig. 7c). The errors of the third type were the misdetected landmarks that occurred when the method accepted noise regions as facial landmarks (Fig. 7a).

5 Discussion

The feature-based method for detecting facial landmarks from neutral and expressive facial images was designed. The method achieved the average detection rate of 90 percent in extracting all four facial landmarks from both neutral and expressive images. The separate percentages were 94 percent for neutral images and 90 percent for expressive ones. The present results revealed that the choice of the oriented edges as the basic features for composing edge maps of the image ensured the invariance in a certain range for eye detection regardless of variations in facial expression and image size. The regions of the left and right eyes were detected in 99 percent of the cases.

However, detecting landmarks of the lower face was affected by changes in expression and image size. The expressions of happiness and disgust had a marked deteriorating effect on detecting the regions of the nose and mouth. The decrease of image size also affected the detection of these landmarks. Variations in expression and decrease in image size attenuated the average detection rates of the mouth and nose regions to 86 percent and of 78 percent, respectively.

The results showed that a majority of errors in detecting facial landmarks occurred at the stage of feature map construction. On the one hand, the results revealed that, often, the nose region remained undetected after the procedure of edge extraction. One possible reason for that was a low contrast of nose regions on the images. As a result, the number of edges extracted from the nose regions was smaller than those extracted from the regions of other landmarks. On the other hand, the threshold limiting number of edges was elaborated for detecting all four facial landmarks. Possibly for this reason, the nose region consisting of a small number of edges, remained undetected.

Another reason for errors in the detection of the nose as well as the mouth was the decrease in image size. The decrease in image size did not affect the contrast around the eyes, but it reduced the contrast around the nose and mouth. Therefore, the number of edges extracted from these regions was reduced and they became less than the threshold and, finally, the nose and mouth regions remained undetected.

On the other hand, the procedure of grouping edges into candidates produced incorrect grouping of several landmarks into the one region. Many errors in constructing regions of the nose and mouth were caused by the use of a fixed neighborhood distance for edge grouping. Utilizing fixed threshold produced a good landmark separation for almost all expressive images (i.e., the error rate in landmark grouping was less than 1 percent). However, the images of happiness and disgust produced a lot of errors in landmark grouping (i.e., the error rates were about 2 percent and 5 percent, respectively). This means that such a fixed neighborhood distance cannot be applied for separating regions of nose and mouth from the happiness and disgust images.

Why were the expressions of happiness and disgust especially difficult to process by the developed algorithms? Probably, the reasons for that were the specific changes of facial appearance while displaying these expressions. There are different AUs and their combinations that are activated during happiness and disgust. In particular, when a face is modified by the expression of happiness, the AU12 is activated. This AU pulled the lips back and obliquely upward.

Further, many of the prototypical disgust expressions suggested by Ekman and Friesen [6] include the activation of AU10. The AU10 lifts the center of the upper lip upward, making the shape of the

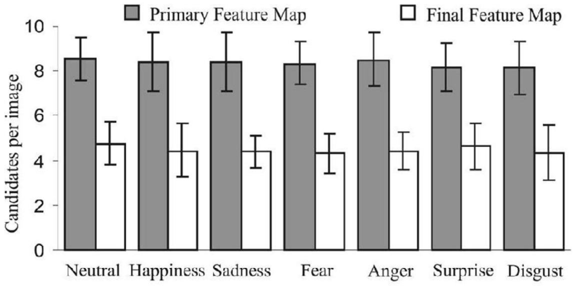


Fig. 6. Average number of candidates per image before and after the procedure of orientation matching. The error bars show plus/minus one standard deviation from the mean values.

TABLE 1
Rate (%) of the Landmark Detection Averaged over Expression and Image Size

Image type		Facial landmark				Average	
		Left eye	Right eye	Mouth	Nose		
Neutral		100	100	91	86		94
Expressive	Happiness	100	100	82	74	89	90
	Sadness	96	96	98	88	95	
	Fear	100	100	93	82	94	
	Anger	100	100	92	84	94	
	Surprise	100	100	98	78	94	
	Disgust	96	96	49	51	73	
Average		99	99	86	78		90

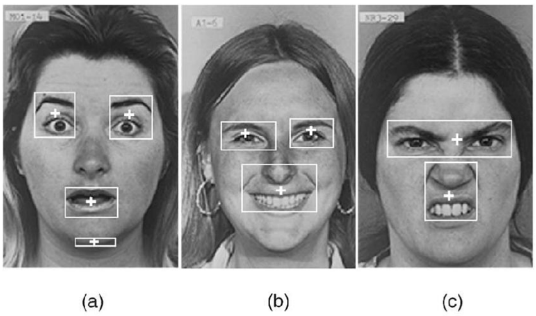


Fig. 7. Errors in detection of facial landmarks. Images from Pictures of Facial Affect. Copyright © 1976 by Paul Ekman. Reprinted with permission of Paul Ekman.

mouth resemble an upside down curve. Both AU10 and AU12 result in deepening the nasolabial furrow and pulling it laterally upward.

Although, there are marked differences in the shape of the nasolabial deepening and mouth shaping for these two AUs, it can be summed up that both expressions of happiness and disgust make the gap between nose and mouth smaller. Such modifications in facial appearance had a marked deteriorating effect on detecting landmarks from the lower part of a face. The neighborhood distances between edges extracted from the regions of nose and mouth became smaller than a threshold. For this reason, edges were grouped together, resulting in incorrect grouping of the nose and mouth regions. The expressions of disgust and sadness (i.e., the combination AU1 and AU4) caused the regions of eyebrows to draw up together, resulting in incorrect grouping regions of both eyes.

One possible way to eliminate errors in landmark separation could be a precise analysis of the density of the edges inside the detected edge regions. The areas with poor point density might contain different areas of edge concentration and could be processed further with some more effective methods like, for example, the neighborhood method.

At the stage of orientation matching, there were some errors in classification between the landmark and noise regions. Although the orientation model revealed a high classification rate for both eyes, it produced errors in classifying the nose region. Such errors were caused by mismatching orientation portraits of the detected candidates and the orientation model. For example, in some cases, the nose region did not have well-defined horizontal dominants in edge orientations—all edge orientations were presented in nearly equal number. Therefore, such a region was rejected as a candidate for facial landmark. On the other hand, errors were caused by the fact that orientation portraits of some noise regions matched the orientation model. In this case, the noise regions were detected. However, most of errors in landmark detection were brought about by errors in the previous stage of feature map construction.

Based on the findings described above, we can conclude that more accurate nose and mouth detection could be achieved by finding some adaptive thresholds for constructing landmark candidates. The overall detection performance of the algorithms could be improved significantly by analyzing spatial configuration of the detected facial landmarks. The use of spatial constraints might be also utilized to predict the location of the undetected facial landmarks [18].

In summary, the method localized facial landmarks with an acceptably high detection rate without a combinatorial increase of complexity of the image processing algorithms. The detection rate of the method was comparable to the detection rate of the known feature-based [15], [17] and color-based [19] methods that have detection rates from 85 to 95 percent, but lower than neural network-based methods [20] with a detection rate of about 96-99.5 percent. Emphasizing the simplicity of the algorithms developed for landmark detection, we conclude they might be implemented as a part of the systems for face and/or facial expression recognition. The discovered errors provided several guidelines for further improvement of the developed method. In our future work, we will focus on finding expression-invariant and robust representations for facial landmarks. Careful attention will be paid to the development of algorithms that are able to cope with images displaying happiness and disgust as the most demanding to process.

ACKNOWLEDGMENTS

This work was financially supported by the Finnish Academy (project number 177857), the Finnish Centre for International Mobility (CIMO), the University of Tampere (UTA), and the Tampere Graduate School in Information Science and Engineering (TISE). The authors would like to thank Professor P. Ekman for his permission to reprint the examples of expressive images from the Pictures of Facial Affect database.

REFERENCES

[1] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 84-91, June 1994.

[2] L. Wiskott, J-M. Fellous, N. Kruger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775-779, July 1997.

[3] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, Oct. 1999.

[4] I. Essa and A. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757-763, July 1997.

[5] Y. Yacoob, H.-M. Lam, and L. Davis, "Recognizing Faces Showing Expressions," *Proc. Int'l Workshop Automatic Face- and Gesture-Recognition*, pp. 278-283, June 1995.

[6] P. Ekman and W. Friesen, *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action.* Palo Alto, Calif.: Consulting Psychologists Press, 1978.

[7] P. Ekman, "The Argument and Evidence about Universals in Facial Expressions of Emotion," *Handbook of Social Psychophysiology*, H. Wagner and A. Manstead, eds. pp. 143-164, Lawrence Erlbaum, 1989.

[8] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System (FACS).* Salt Lake City: A Human Face, 2002.

[9] A. Fridlund, "Evolution and Facial Action in Reflex, Social Motive, and Paralanguage," *J. Biological Psychology*, vol. 32, pp. 3-100, Feb. 1991.

[10] V. Surakka and J. Hietanen, "Facial and Emotional Reactions to Duchenne and Non-Duchenne Smiles," *Int'l J. Psychophysiology*, vol. 29, pp. 23-33, June 1998.

[11] E. Hjelmas and B. Low, "Face Detection: A Survey," *J. Computer Vision and Image Understanding*, vol. 83, pp. 235-274, Sept. 2001.

[12] M. Yang, D. Kriegman, and N. Ahuaja, "Detecting Face in Images: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 34-58, June 2002.

[13] Y. Liu, K. Schmidt, J. Cohn, and S. Mitra, "Facial Asymmetry Quantification for Expression Invariant Human Identification," *J. Computer Vision and Image Understanding*, vol. 91, pp. 138-159, Aug. 2003.

[14] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, Feb. 2001.

[15] A. Golovan, "Neurobionic Algorithms of Low-Level Image Processing," *Proc. Second All-Russia Scientific Conf. Neuroinformatics*, vol. 1, pp. 166-173, May 2000.

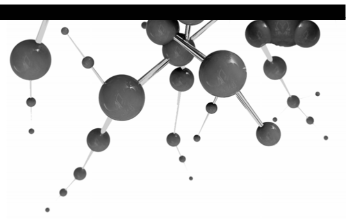
[16] P. Ekman and W. Friesen, *Pictures of Facial Affect.* Palo Alto, Calif.: Consulting Psychologists Press, 1976

[17] D. Shaposhnikov, A. Golovan, L. Podladchikova, N. Shevtsova, X. Gao, V. Gusakova, and I. Guizatdinova, "Application of the Behavioral Model of Vision for Invariant Recognition of Facial and Traffic Sign Images," *J. Neurocomputers: Design and Application*, vol. 7, no. 8, pp. 21-33, 2002.

[18] G. Yang and T. Huang, "Human Face Detection in a Complex Background," *J. Pattern Recognition*, vol. 27, pp. 53-63, Jan. 1994.

[19] K. Sobottka and I. Pitas, "Extraction of Facial Regions and Features Using Color and Shape Information," *Proc. Int'l Conf. Pattern Recognition*, vol. 3, pp. 421-425, Aug. 1996.

[20] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 45-51, June 1998.



Publication II

Guizatdinova I. and Surakka V. (2005). Detection of Facial Landmarks from Neutral, Happy, and Disgust Facial Images. *Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'05)*, UNION Agency: Science Press, 55-62.

Available online at:
http://wscg.zcu.cz

Detection of Facial Landmarks from Neutral, Happy, and Disgust Facial Images

Ioulia Guizatdinova and Veikko Surakka
Research Group for Emotions, Sociality, and Computing
Tampere Unit for Computer-Human Interaction
Department of Computer Sciences
University of Tampere
FIN-33014 Tampere, Finland
ig74400@cs.uta.fi
Veikko.Surakka@uta.fi

ABSTRACT

Automated analysis of faces showing different expressions has been recently studied to improve the quality of human-computer interaction. In this framework, the expression-invariant face segmentation is a crucial step for any vision-based interaction scheme. A method for detecting facial landmarks from neutral and expressive facial images was proposed. In present study, a particular emphasis was given to handling expressions of happiness and disgust. The impact of these expressions on the developed method was tested using dataset including neutral, happiness and disgust images. The results demonstrated a high accuracy in detecting landmarks from neutral images. However, the expressions of happiness and disgust had a deteriorating effect on the landmark detection.

Keywords
Image processing, face segmentation, detection of local oriented edges, Gaussian, facial landmarks, human-computer interaction.

1. INTRODUCTION

In the past decades there has been a considerable interest in improving all aspects of human-computer interaction (HCI). One way to achieve intelligent HCI is making computers to interact with user in the same manner as it takes place in human-human interaction.

Humans naturally interact with each other through verbal (i.e. speech) and nonverbal (i.e. facial expressions, gesture, vocal tones, etc.) sign systems. It is argued that during human-human interaction only a small part of the conveyed messages is verbally communicated, and the greatest part is nonverbally coded. Considering nonverbal communication, it is possible to say that facial expressions occupy about a half of the transmitted signals. In the context of user-friendly HCI, a face is an important source of information about the user to be analyzed by the computer.

WSCG 2005 conference proceedings, ISBN 80-903100-7-9
WSCG'2005, January 31-February 4, 2005
Plzen, Czech Republic.
Copyright UNION Agency – Science Press

Automated analysis of a computer user's face has recently become an active research field in the computer vision community. Different vision-based schemes for intelligent HCI are currently being developed. The ability of a computer to detect, analyse and, finally, recognize a user's face has many applications in the domain of HCI.

The analysis and recognition of facial expressions in the context of HCI are elements of interaction design called affective computing [Jen98]. The main idea of the affective computing is that the computer detects the user's affective state and takes an appropriate action, for example, offers assistance for the user or adapts to the user's needs. Proper detection of the changes in the user's facial cues is a precondition for the computer to take any emotionally or otherwise intelligent socially interactive actions towards the user.

The Facial Action Coding System (FACS) [Ekm78] is widely used to analyse visually observable facial expressions. FACS has been developed for objective analysis of any changes in the facial appearances.

According to the FACS, a muscular activity producing changes in facial appearance is coded in the terms of action units (AU). Certain specific combinations of AUs have been frequently suggested to represent seven prototypical facial displays: neutral, happiness, sadness, fear, anger, surprise, and disgust.

It is known that reliable person identification and verification are important cornerstones for improving security in various contexts of information society. A natural means of identifying person that gives a close resemblance to the way how humans recognize persons is analysing a person's face.

Face identification has two important advantages. First, it requires a minimal interaction with a person, for example, compared with such biometrics as prompted speech or fingerprints. Second, it is impossible to lose or forget a face as it might happen with passwords or key-cards.

In this framework, automated detection of a face and its features is considered to be an essential requirement for any vision-based HCI scheme [Don99, Wis97]. However, due to such factors as illumination, head pose, expression and scale, facial features vary greatly in their appearance. It is shown that facial expressions are particularly important factors affecting the automated detection of facial features [Yac95]. Nowadays the problem of effective and expression-invariant face detection and segmentation still remains unsolved.

In our previous study we have proposed a method for detecting facial landmarks from neutral and expressive facial images [GuiS]. The developed approach has combined a feature-based method for face segmentation [Sha02] and a profound knowledge on how different facial muscle activations modify the appearance of a face during emotional and social reactions [Par04, Sur98].

Experimented findings have revealed that detection of landmarks from the lower part of a face was especially affected by expressions of happiness and disgust. In particular, detection of the nose and mouth produced the greatest number of detection errors. We assumed that these expressions modify the lower face so that it becomes difficult to differentiate lower face landmarks like nose and mouth. For this reason the present aim was to analyse an accuracy of landmark detection from images of happiness and disgust to corroborate the previous findings.

2. FACIAL LANDMARK DETECTION

The method for detection of facial landmarks consisted of three stages: image preprocessing, image map constructing and orientation matching [GuiS]. These stages are described below.

2.1. Image preprocessing

First, an image was transformed into the 256-grey-level-scale format. Then, a recursive Gaussian transformation was used to smooth the grey-level image [Gol00]. Image smoothing reduced a search space for detecting facial features (i.e. eliminated noise edges and removed small details) [Can86].

In the following stages of the landmark detection, the smoothed grey-level images were used to detect candidates for facial landmarks. The non-smoothed grey-level images allowed us to analyse the detected candidates in details. In that way, the amount of information to be processed was significantly reduced.

2.2. Image map constructing

The local high-contrast oriented edges were used as basic features for constructing edge maps of the image [Ryb98]. Apart from previous studies [Sha02], we decreased a number of edge orientations to construct edge maps of the image. In particular, we used $2 \div 6$ and $10 \div 14$ edge orientations (see Fig.1). Decreasing a number of edge orientations allowed us to reduce sufficiently the computational complexity of the method.

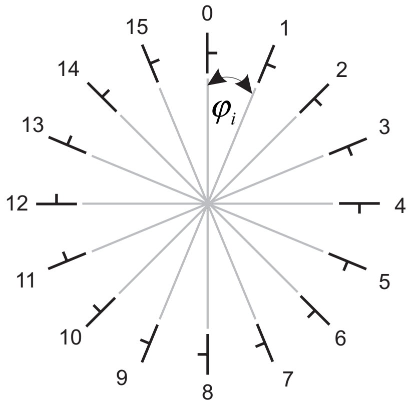


Figure 1. Orientation template, $\varphi_i = i \cdot 22.5°$ **,** $i = 0 \div 15$ **.**

The oriented edges were extracted by convolving the smoothed image with a set of ten convolution kernels. Each kernel was sensitive to one out of ten chosen edge orientations. For each pixel, the contrast magnitude of a local edge was estimated with maximum response of ten kernels at this pixel location. The orientation of a local edge was estimated with orientation of a kernel that gave the maximum response. The whole set of ten kernels resulted from differences between two oriented Gaussians with shifted kernels.

After the local oriented edges had been extracted, they were filtered by a contrast. The threshold for contrast filtering was determined as an average contrast of the whole smoothed image.

Then, the extracted oriented edges were grouped into edge regions presumed to contain facial landmarks. Edge grouping was based on neighbourhood distances between edges and was limited by a number of possible neighbours for each oriented edge. The optimal thresholds for edge grouping were determined using a small set of expressive images of the same person. The optimal thresholds represented landmark candidates as regions of connected edges that were well separated from the rest of edges.

Once the limits of edge regions had been detected, these regions were analysed more precisely. The procedures of edge extracting, contrast thresholding and edge grouping were applied to the non-smoothed image within the limits of the extracted edge regions. The threshold for contrast filtering was determined as a double average contrast of the non-smoothed image.

In the end, the primary image map consisted of edge regions representing candidates for facial landmarks. The centres of mass determined the locations of the landmark candidates. In the next stage, the landmark candidates were analysed according to their orientation description and matched with an orientation model.

2.3. Orientation matching

The orientation portraits of the landmark candidates were constructed on the basis of their local orientation description. The analysis of the orientation portraits revealed four important findings.

First, local oriented edges extracted within regions of eyebrows, eyes, nose and mouth had a characteristic density distribution. Thus, the orientation portraits of these landmarks had two dominant horizontal orientations. The results of the present study corroborated our previous findings [Sha02].

Second, we found that prototypical facial expressions did not affect the distribution of the oriented edges in the regions of facial landmarks [GuiS]. The orientation portraits of facial landmarks still had the same structure including two dominants corresponding to horizontal orientations (see Appendix 1*a*).

Moreover, for the regions of eyes and mouth the number of edges corresponding to horizontal orientations was more than 50% larger when compared to a number of edges corresponding to other orientations. All edge orientations were represented by non-zero number of the edges.

Third, the average orientation portraits of facial landmarks revealed the same structure including two horizontal dominants (see Fig.2, Appendix 2) [GuiS].

Fourth, noise regions extracted from the expressive images had an arbitrary distribution of the oriented

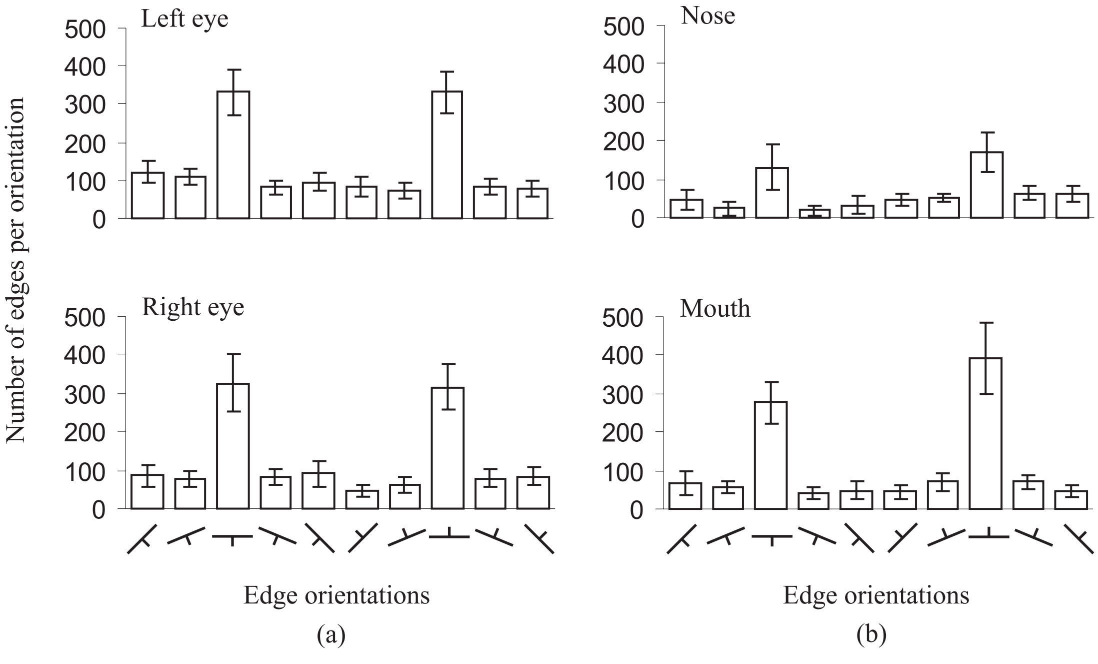


Figure 2. Orientation portraits of facial landmarks averaged over prototypical facial displays.

edges and often had orientations represented by zero number of edges (see Appendix 1*b*).

The knowledge on clear-cut distinction between orientation portraits of facial landmarks and noise regions allowed us to verify the existence of a landmark on the image. To do that, the orientation portraits of facial candidates were matched with an orientation model of facial landmarks.

2.3.1. Orientation model

The characteristic orientation model for detecting facial landmarks consisted of ten possible edge orientations, namely, edge orientations ranging from $45°$ to $135°$ and $225°$ to $315°$ in step of $22.5°$.

The following rules defined the structure of the orientation model: (a) horizontal orientations are represented by the biggest number of edge points; (b) a number of edges corresponding to each of the horizontal orientations is more than 50% bigger than a number of edges corresponding to other orientations taken separately; and (c) orientations can not be represented by zero-number of edge points.

The candidates that did not correspond to the orientation model were removed from the final image map. In such a way, the procedure of orientation matching filtered the regions containing landmarks from the noise.

The detected candidates for facial landmarks were further classified manually into one of the following groups: noise or facial landmark (i.e. eye-eyebrow, nose and mouth).

3. DATABASES

To evaluate the accuracy of the proposed method we used the Pictures of Facial Affect (PFA) database [Ekm76] and the Cohn-Kanade Face (CKF) database [Kan00].

The PFA database consisted of 110 frontal-view images of 14 individuals (i.e. 6 males and 8 females) representing neutral and six prototypical facial expressions of emotions: happiness, sadness, fear, anger, surprise and disgust. On average, there were about sixteen pictures per expression. The size of the images was preset into 250 by 300 pixels.

The CKF images were originally coded using single AUs and their combinations. In according to translation rules defined in the Investigator's Guide to the FACS manual [Ekm00], the images were relabelled into the emotional prototypes. The images corresponding to the prototypes of happiness and disgust were selected. Thus, there were 172 images: 65 neutral images, 65 images of happiness and 42 images of disgust expression. All the images were normalized to contain only a facial part of the original image. Either of the datasets included faces with facial hair and glasses. All the images were resized into 250 by 480 pixel arrays.

The PFA database was used to select the optimal thresholds for edge grouping and to construct the landmark orientation model [GuiS]. In present study, the CKF database was used to test the accuracy of the method in detection of facial landmarks specifically from the images showing happiness and disgust.

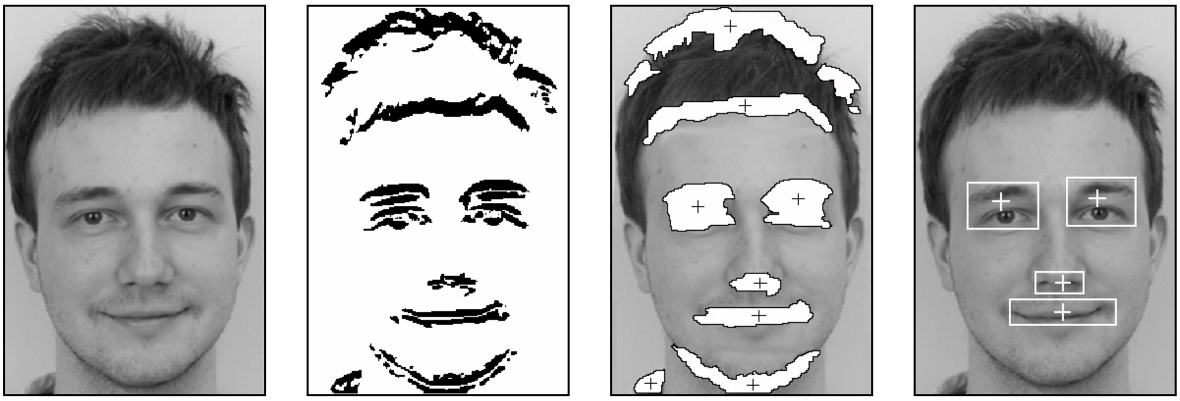

(a) (b) (c) (d)

Figure 3. (a) original facial image; (b) extracted local oriented edges (black dots); (c) primal edge map represents candidates for facial landmarks (white regions) and their mass centers (crosses); (d) final edge map represents the detected facial landmarks.

4. RESULTS

Figure 3 gives an example of edge map composed of the local oriented edges extracted from the expressive facial images. Thus, local edges of $45° \div 135°$ and $225° \div 315°$ defined in step of $22.5°$ constituted the edge map of the happy image shown on Figure 3*b*. Figure 3*c* demonstrates the edge map after contrast thresholding and grouping extracted edge points into the candidates for facial landmarks. Figure 3*d* illustrates the final image map that included only the candidates having orientation portraits well matched with the orientation model.

The average number of the candidates per image of the primary edge map was 7.46. The results revealed that variations in facial expressions did not affect significantly the average number of the candidates per image. The average number of candidates per image was reduced to 3.71 for the final edge map. Such a fact allows us to claim that the procedure of orientation matching reduced the number of landmark candidates by 50%. Figure 4 illustrates the decrease in the number of candidates per image averaged over neutral, happy, and disgust images.

The accuracy of the proposed method was calculated as a ratio of the number of detected landmarks to the number of images used in testing. As it can be seen from Table 1, the developed method achieved a sufficiently high accuracy of 95% in detecting all four facial landmarks from the neutral images. As it can be seen from the table, both eyes are represented as a single column since these landmarks had equal detection accuracy.

However, the results showed that the expressions of happiness and disgust had a marked deteriorating effect on detecting facial landmarks. It is noteworthy that the detection of nose and mouth was more affected by facial expressions than the detection of eyes.

Three types of detection errors caused the decrease in detection accuracy. Figure 5 gives examples of such errors. The undetected facial landmarks were considered to be the errors of the first type. Such errors occurred when a facial landmark was rejected as a noise region after orientation matching. In particular, the nose was the most undetectable facial landmark (see Fig. 5*a*).The incorrectly grouped landmarks were regarded as the errors of the second type. The most common error of the second type was grouping regions of nose and mouth into one region (see Fig. 5*b*). The errors of the third type were the misdetected landmarks that occurred when the noise regions were accepted as the facial landmarks (see Fig. 5*c*).

	Eye	**Nose**	**Mouth**	**Average**
Neutral	98	92	92	95
Happiness	100	50	50	75
Disgust	67	57	59	62
Neutral & Expressive	88	66	67	78

Table 1. Average accuracy (%) of the landmark detection

5. CONCLUSIONS

The method for detecting facial landmarks from both neutral and expressive facial images was presented and described. The method revealed an average accuracy of 95% in detecting four facial landmarks from neutral facial images.

However, the detection of facial landmarks from happy and disgust facial images produced a large number of detection errors. Thus, the expressions of happiness and disgust attenuated the average (i.e. over all regions) detection accuracy to 75% and of 62%, respectively. Especially the detection of nose and mouth were affected by both expressions of disgust and happiness. These expressions deteriorated the detection of nose and mouth to 50% for happiness. For the disgust expression the detection of nose and mouth deteriorated to 57 and 59, respectively. The present results corroborated our earlier findings that facial expressions have a marked deteriorating effect on the landmark detection

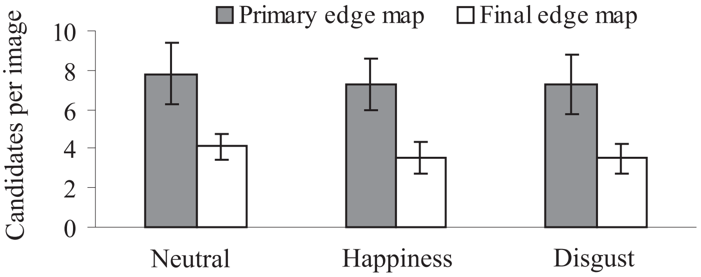


Figure 4. Average number of candidates per image before and after orientation matching.

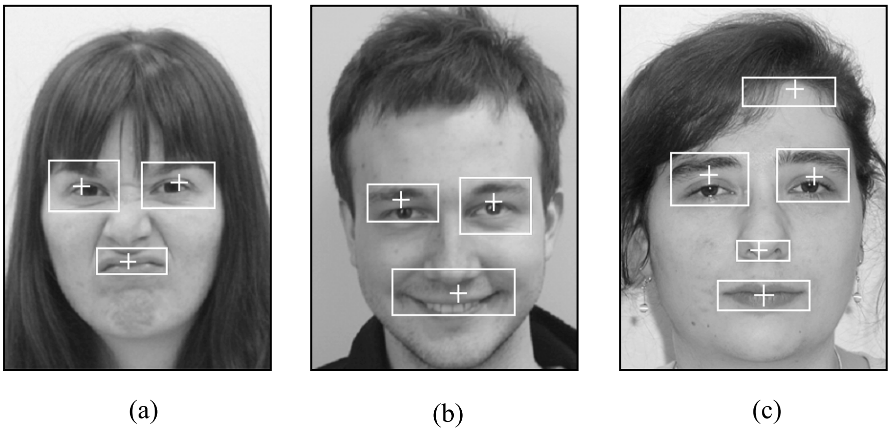

(a) (b) (c)

Figure 5. Examples of the detection errors: (a) undetected nose; (b) incorrectly grouped nose and mouth; (c) detected noise region.

algorithms.

In summary, the accuracy of the landmark detection from neutral images was comparable with a detection accuracy of the known feature-based and colour-based methods though it is lower than neural network-based methods. The algorithms developed for landmark detection were simple and fast enough to be implemented as a part of systems for face and/or facial expression recognition.

The detection of facial landmarks from expressive images, especially from happy and disgust images needs to be improved. This is especially important in order to make a computer differentiate between positive expressions of emotions, for example, smiling and some negative expressions like disgust. To detect and differentiate between positive and negative user emotions, it is the very minimum prerequisite for affective HCI. This kind of an improvement of the method is also a precondition for recognizing facial identity of a user as well.

6. ACKNOWLEDGMENTS

This work was financially supported by the Finnish Academy (project number 177857), the Finnish Centre for International Mobility (CIMO), the University of Tampere and the University of Tampere Foundation.

7. REFERENCES

[Can86] Canny, J. A computational approach to edge detection. IEEE Trans. on Pattern Analysis and Machine Intelligent 8, No.6, pp.679–98, 1986.

[Don99] Donato, G., Bartlett, M., Hager, J., Ekman, P., and Sejnowski, T. Classifying facial actions. IEEE Trans. on Pattern Analysis and Machine Intelligent 21, No.10, pp.974–989, 1999.

[Ekm76] Ekman, P., and Friesen, W. Pictures of facial affect. Consulting Psychologists Press, Palo Alto, California, 1976.

[Ekm78] Ekman, P., and Friesen, W. V. Facial Action Coding System (FACS): A technique for the measurement of facial action. Consulting Psychologists Press, Palo Alto, California, 1978.

[Ekm00] Ekman, P., Friesen, W., and Hager, J. Facial Action Coding System (FACS). UTAH: A Human Face, Salt Lake City, 2002.

[Gol00] Golovan, A. Neurobionic algorithms of low-level image processing. in Second All-Russia Scientific Conference Neuroinformatics-2000 conf.proc., vol. 1, pp.166-173, 2000.

[GuiS] Guizatdinova, I., and Surakka, V. Detection of facial landmarks from emotionally expressive and neutral facial images. IEEE Trans. on Pattern Analysis and Machine Intelligent, submitted.

[Jen98] Jennifer, H., and Picard, J. Digital processing of affective signals. in IEEE ICASSP'98 conf.proc., Seattle, 1998.

[Kan00] Kanade, T., Cohn, J.F., and Tian, Y. Comprehensive database for facial expression analysis. in AFGR'00 conf.proc., Grenoble, p.46, 2000.

[Par04] Partala, T., and Surakka, V. The effects of affective interventions in human-computer interaction. Interacting with Computers, 16, pp.295-309, 2004.

[Ryb98] Rybak, I. A model of attention-guided invariant visual recognition. Vision Research 38, No.15/16, pp.2387-2400, 1998.

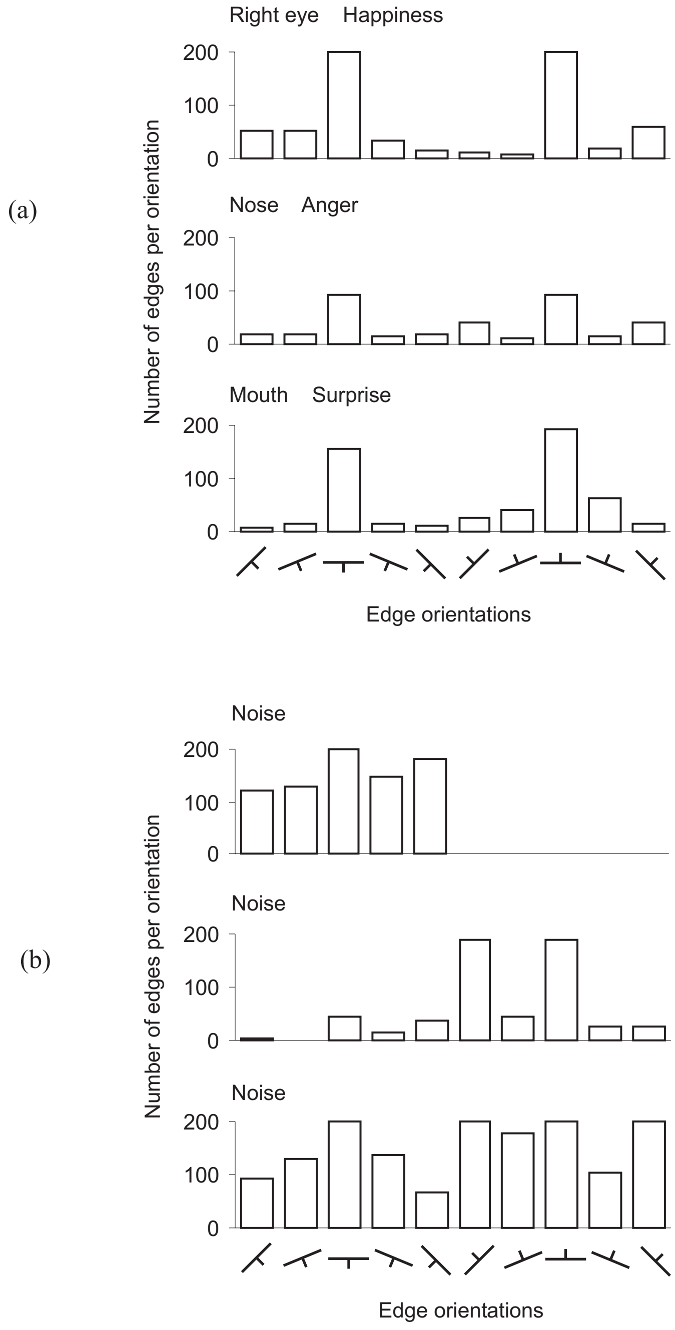
[Sha02] Shaposhnikov, D., Golovan, A., Podladchikova, L., Shevtsova, N., Gao, X., Gusakova, V., and Gizatdinova, Y. Application of the behavioural model of vision for invariant recognition of facial and traffic sign images. Neurocomputers: Design and Application 7, No.8, pp.21-33, 2002.

[Sur98] Surakka, V., and Hietanen, J. Facial and emotional reactions to Duchenne and non-Duchenne smiles. International Journal of Psychophysiology 29, pp.23-33, 1998.

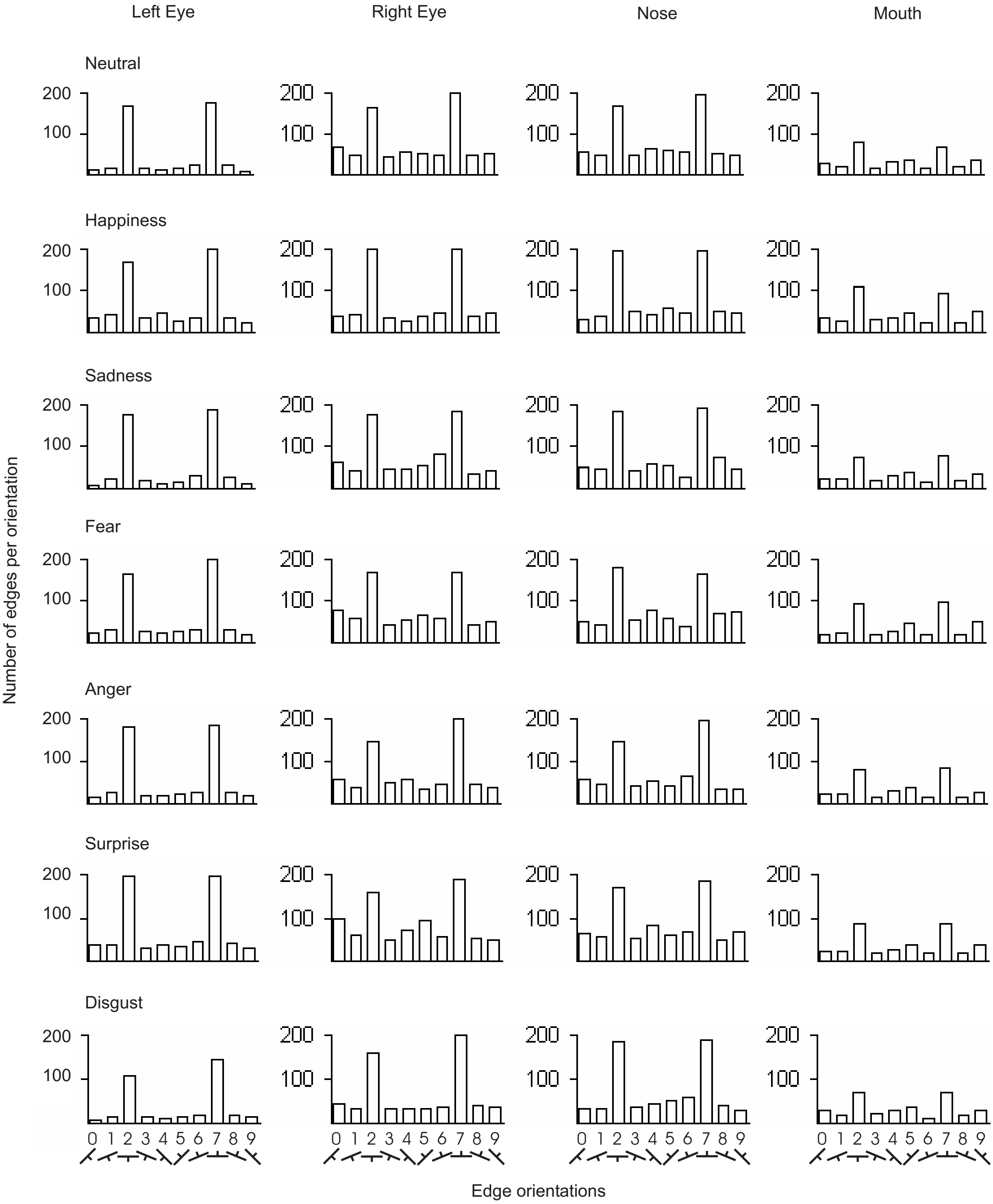
[Wis97] Wiskott, L., Fellous, J-M., Kruger, N., and von der Malsburg, C. Face recognition by elastic bunch graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligent 19, No.7, pp.775-779, 1997.

[Yac95] Yacoob, Y., Lam, H-M., and Davis, L. Recognizing faces showing expressions. in IWAFGR'95 conf.proc., Zurich, pp.278-283, 1995.

Appendix 1. Orientation portraits of (a) landmarks with characteristic edge distribution, and (b) noise regions with arbitrary edge distribution.



Appendix 2. Average orientation portraits for facial landmarks. The columns represent four facial landmarks and rows represent seven prototypical facial displays.



Publication III

Gizatdinova Y. and Surakka V. (2008). Effect of Facial Expressions on Feature-Based Landmark Localization in Static Grey Scale Images. *Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP'08)*, INSTICC, 259-266.

Abstract is available online at:
http://www.visapp.org

EFFECT OF FACIAL EXPRESSIONS ON FEATURE-BASED LANDMARK LOCALIZATION IN STATIC GREY SCALE IMAGES

Yulia Gizatdinova and Veikko Surakka
Research Group for Emotions, Sociality, and Computing, Tampere Unit for Computer-Human Interaction (TAUCHI)
University of Tampere, Kanslerinnrinne 1, 33014, Tampere, Finland
{yulia.gizatdinova, veikko.surakka}@cs.uta.fi

Keywords: Image processing and computer vision, segmentation, edge detection, facial landmark localization, facial expressions, action units.

Abstract: The present aim was to examine the effect of facial expressions on the feature-based landmark localization in static grey scale images. In the method, local oriented edges were extracted and edge maps of the image were constructed at two levels of resolution. Regions of connected edges represented landmark candidates and were further verified by matching against the edge orientation model. The method was tested on a large database of expressive faces coded in terms of action units. Action units described single and conjoint facial muscle activations in upper and lower face. As results demonstrated, eye regions were located with high rates in both neutral and expressive datasets. Nose and mouth localization was more attenuated by variations in facial expressions. The present results specified some of the critical facial behaviours that should be taken into consideration while improving automatic landmark detectors which rely on the low-level edge and intensity information.

1 INTRODUCTION

Facial expressions result from contractions and/or relaxations of facial muscles. These non-rigid facial movements result in considerable changes of facial landmark shapes and their location on the face, presence/absence of teeth, out-of-plan changes (showing the tongue), and self-occlusions (bitted lips). The best known and most commonly referred linguistic description of facial expressions is the Facial Action Coding System (FACS) (Ekman and Friesen, 1978; Ekman, Friesen, and Hager, 2002). The FACS codes an expressive face in terms of action units (AUs). The numerical AU code describes single and conjoint facial muscle activations. It is anatomically-based and therefore represents facial expressions as a result of muscle activity without referring to emotional or otherwise cognitive state of a person on the image.

It was suggested that structural changes in the regions of facial landmarks (eyebrows, eyes, nose, and mouth) are important and in many cases sufficient for AU recognition. In automatic AU recognition, manual preprocessing is typically needed to select a set of fiducial points (for example, eye centres and mouth corners) in static image or initial frame of the video sequence. Fiducial points are further used to track changes in the face resulted from its expressive behaviour or to align an input image with a standard face model. Currently, there is a need for a system that can automatically locate facial landmarks in the image prior to the following steps of the automatic facial expression analysis.

In static facial image, there is no temporal information on facial movements available. Facial landmark localization in this case is generally addressed by modelling a local texture in the regions of landmarks and by modelling a spatial arrangement of the found landmark candidates (Hjelmas and Low, 2001; Pantic and Rothkrantz, 2000; Yang, Kriegman, and Ahuaja, 2002). The main challenge is to find a representation of the landmarks that efficiently characterizes a face and remains robust with respect to facial deformations brought about by facial expressions.

Addressing the problem of expression invariant localization of facial landmarks in static grey scale images, the feature-based method was introduced (Gizatdinova and Surakka, 2006). In the method,

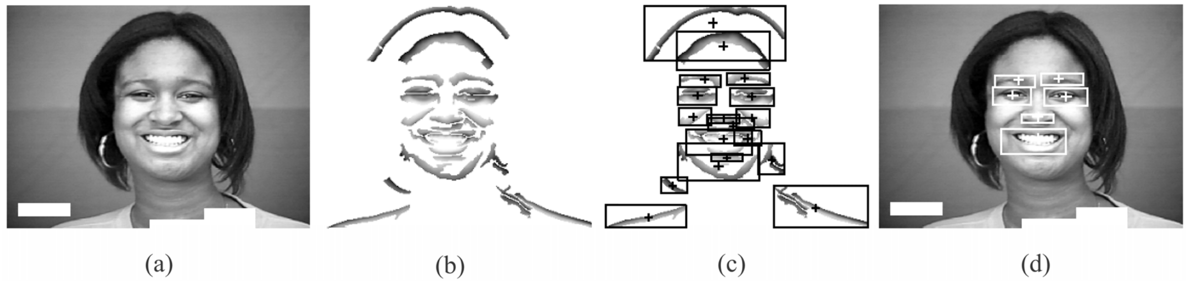


Figure 1: Facial landmark localization: (a) original image, (b) parts of the image located as regions of connected edges; (c) landmark candidates; (d) final localization result after edge orientation matching. Bounding boxes indicate locations and crosses define mass centres of the found regions. Image indexes are masked by white boxes. Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression Database (Kanade, Cohn, and Tian, 2000). Reprinted with permission.

edge representation of the face was taken at ten edge orientations and two resolution levels to locate regions of eyes (including eyebrows), lower nose, and mouth. The resulted edge map of the image consisted of regions of connected local oriented edges presumed to contain facial landmarks. To verify the existence of a landmark on the image, the extracted landmark candidates were matched against the edge orientation model. Figure 1 illustrates the main steps of the method. The description of edge detection, edge grouping, and edge orientation matching steps is given in more detail in Appendixes A and B.

A degradation in the landmark localization rates was reported for expressive dataset as compared to neutral dataset. The further analysis (Guizatdinova and Surakka, 2005) suggested that there were certain AUs which significantly deteriorated the performance of the method. It was assumed that AUs activated during happiness (AU12), disgust (AU 9 and 10), and sadness (AU 1 and 4) would be such central AUs. Having such a ground, the main motivation for the present study was the fact that although a degradation in the landmark localization rates due to expression variations is generally appreciated in the computer vision society; however, a little attempt has been done to analyze what muscle activations cause the degradation. To estimate more accurately what facial muscular activity affects the feature-based landmark localization, a more detailed study was needed.

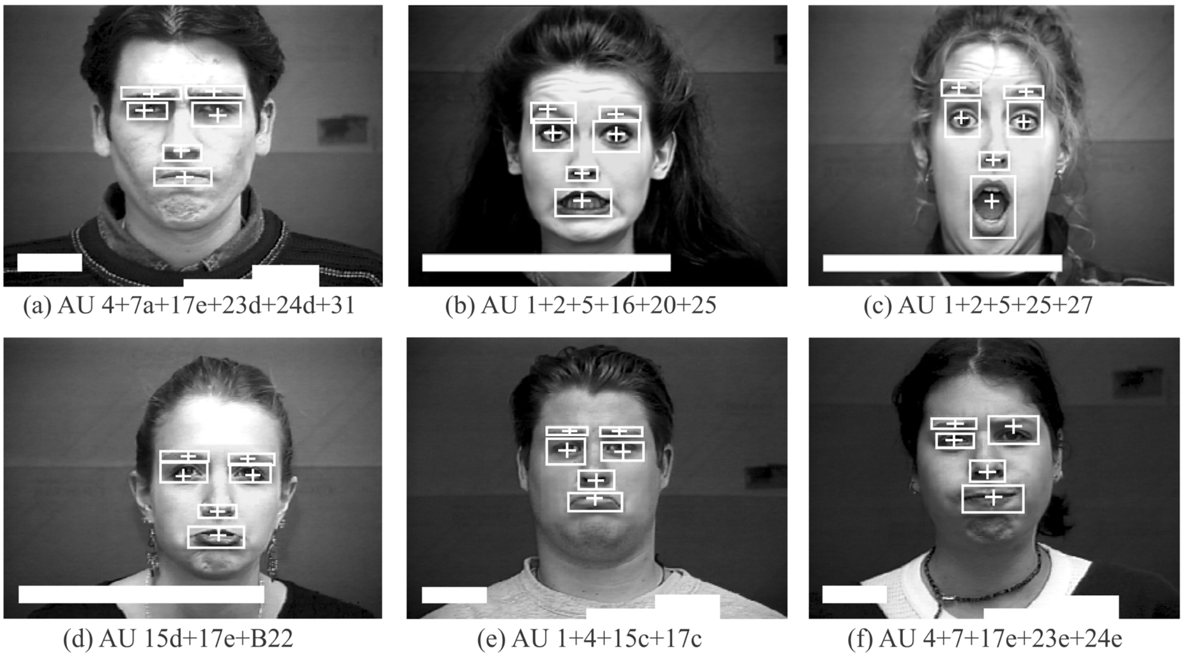
The present aim was to evaluate the developed method on a larger AU-coded database of expressive images and investigate the impact of single AUs and AU combinations on the facial landmark localization in static facial images.

2 DATABASE

The Cohn-Kanade AU-Coded Facial Expression Database (Kanade, Cohn, and Tian, 2000) was used to test the method. The database consists of image sequences taken from 97 subjects of both gender (65% female) with ages varying from 18 to 30 years. The database represents subjects with different ethnic background (81% Caucasian, 13% African-American, and 6% Asian or Latino). There were no images with eye glasses and strong facial hair.

Each image sequence starts with a neutral face that gradually transforms to an expressive one. Expressions from different sequences can differ in levels of intensity. Expressive images are labelled in terms of AUs, and AUs occur both alone and in combinations. The AU descriptors taken from the FACS manual (Ekman, Friesen, and Hager, 2002) are as follows. Upper face AUs: 1 - inner eyebrow raiser, 2 - outer eyebrow raiser, 4 - eyebrow lowerer, 5 - upper lid raiser, 6 - cheek raiser and lid compressor, 7 - lid tightener, 43 - eye closure, and 45 - blink. Lower face AUs: 9 - nose wrinkler, 10 - upper lip raiser, 11 - chin raiser, 12 - lip corner depressor, 14 - lips part, 15 - jaw drop, 16 - mouth stretch, 17 - lower lip depressor, 18 - lip pucker, 20 - lip tightener, 23 - lip presser, 24 - nasolabial furrow deepener, 25 - lip corner puller, 26 - lip stretcher, and 27 – dimpler.

From each image sequence, the first and the last frames were selected which corresponded to neutral and expressive faces, respectively. A total of 468 neutral and 468 expressive images were selected. All images were scaled to approximately 300 by 230 pixel arrays. No face alignment was performed. Image indexes were masked by white boxes.



(a) AU 4+7a+17e+23d+24d+31

(b) AU 1+2+5+16+20+25

(c) AU 1+2+5+25+27

(d) AU 15d+17e+B22

(e) AU 1+4+15c+17c

(f) AU 4+7+17e+23e+24e

Figure 2: Examples of correctly located facial landmarks. Bounding boxes indicate locations and crosses define mass centres of the found regions. Image indexes are masked by white boxes. Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression Database (Kanade, Cohn, and Tian, 2000). Reprinted with permission.

3 LANDMARK LOCALIZATION

All the localization results were checked manually and classified into one of the following groups: correct, wrong, and false localization. Different from systems in which a point defines the localization result, in this study the localization result was defined as a rectangular bounding box placed over the located region. The mass centre of the located region indicated an estimate of the centre of the landmark.

A correct landmark localization was considered if the bounding box overlapped approximately more than a half of the visible landmark and enclosed the area surrounding a landmark less than the actual area of the landmark (Figure 2). Eye localization was counted correct if the bounding box included both eye and eyebrow, or eye and eyebrow were located separately. In case if eyebrow was located as a separate region, it was obligatory that a corresponding eye was also found.

A wrong landmark localization was considered if the bounding box covered several neighbouring facial landmarks. Wrong landmark localization was observed in 0.54 cases per image. For this type of localization error, the failure in nose and mouth localization was mainly due to the effect of lower face AUs 9, 10 and 12. These AUs, occurring alone or in combinations, produced the erroneous grouping of nose and mouth into one region. AUs 4, 6, 7, and their combinations with other AUs sometimes caused the merging of the eye regions.

A false landmark localization was considered if the bounding box included some non-landmark regions as, for example, elements of clothing, hair or face parts like wrinkles, shadows, ears, and eyebrows located without a corresponding eye. The procedure of orientation matching reduced the average number of candidates per image into almost a half for neutral (from 6.57 to 3.49) and expressive (from 6.97 to 3.60) images, see Figure 3,a. Accordingly, the average number of false localizations per image was reduced from 1.84 to 0.01 for neutral and from 2.07 to 0.08 expressive images, see Figure 3,b. Figure 4 shows some examples of the localization errors.

Table 1 summarizes the performance of the method. For each landmark, a rate of its localization was defined as a ratio between the total number of correctly located landmarks and the total number of images used in testing (as there was one landmark per image). A false positive was defined as a number of false localizations.

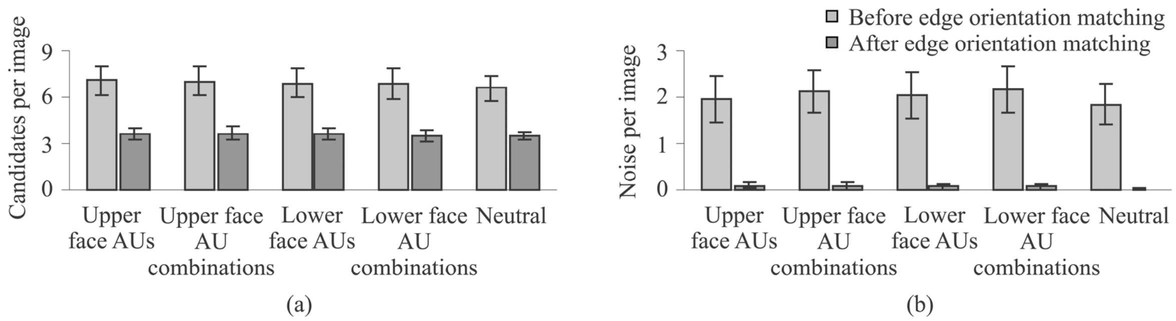
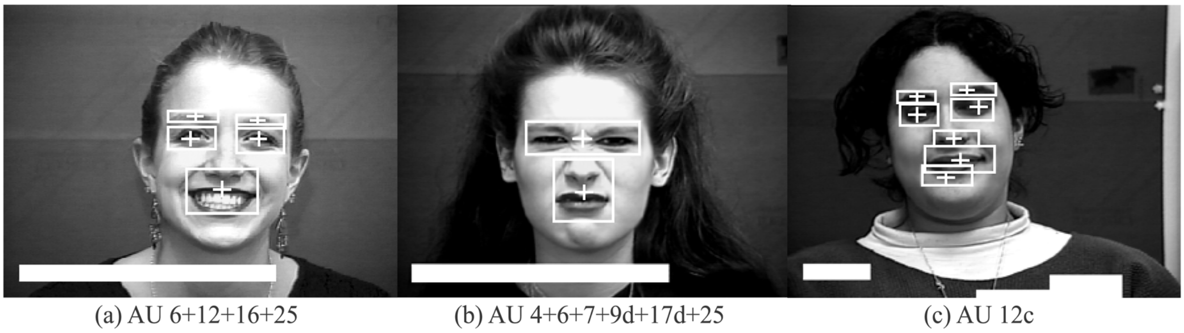


Figure 3: Average number of landmark candidates per image before and after the procedure of orientation matching. The error bars show plus/minus one standard deviation from the mean values.



(a) AU 6+12+16+25 (b) AU 4+6+7+9d+17d+25 (c) AU 12c

Figure 4: Examples of errors in facial landmark localization: (a) nose and mouth wrong localization; (b) eye region wrong localization and nose and mouth wrong localization; (c) false localization. Bounding boxes indicate locations and crosses define mass centres of the found regions. Image indexes are masked by white boxes. Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression Database (Kanade, Cohn, and Tian, 2000). Reprinted with permission.

Table 1: Performance of the method on neutral and expressive datasets.

Dataset	Rates of landmark localization				Total	False positive
	R eye region	L eye region	Nose	Mouth		
Neutral	98%	99%	93%	91%	95%	9
Expressive	93%	93%	55%	55%	74%	55

The method achieved average localization rate of 84% in finding all facial landmarks. On the whole, localization rates were better for neutral than for expressive images. Thus, eye regions were located with high rates in both neutral and expressive datasets. However, nose and mouth localization rates were considerably better for neutral than for expressive images. In the next sections, the effect of single AUs and AU combinations on the landmark localization rates will be considered.

3.1 Effect of Facial Expressions on Landmark Localization Rates

The results of the previous section demonstrated the degradation of the landmark localization rates in case of expressive dataset. The same results can be interpreted in a way that specifies what facial behaviours caused the degradation. At this point we aimed to analyze the effect of upper and lower face AUs on the landmark localization rates. To do that the localization results were classified systematically using the following approach. The results were combined into four AU groups according to AUs presented in the test image, see Table 2. Thus, if image label included single AU, the localization result was classified into group I or II. If image label included a combination of two AUs, the localization result was classified into group III or IV. AU43 (eye closure) and AU45 (blink) were combined together because they both have the same visual effect on the facial appearance and different durations of these AUs can not be measured from the static images.

Due to the fact that some AUs were not presented in the database or the number of images was too few (less than 6), only a limited number of AUs and AU combinations was used. The classification allowed the results to belong to more than one group. On the next step, average landmark localization rates were calculated for each AU subgroup. Tables 3 and 4 illustrate the effect of chosen AU groups on the landmark localization rates. In the tables, AUs and AU combinations were defined as having no or slight effect if average localization rates were in the range of 90-100%, as medium if localization rates were in the range of 80-89%, and strong if localization rates were below 79%. Table 3 demonstrates that eye region localization was consistently good in the context of the presented AU groups. Among all the facial behaviours, upper face AU9 and AU combinations 4+6, 9+25, and 10+17 had the most deteriorating effect on the eye region localization. Lower face AU 9 and AU combinations 4+6, 9+17, 12+20, 12+16 had the most deteriorating effect on the nose and mouth localization in Table 4. In the tables, bold font defines AUs and AU combinations which had the strongest effect on both upper and lower face landmark localization.

4 DISCUSSION

The effect of facial expressions on the feature-based localization of facial landmarks in static facial images was evaluated. In this section, the impact of upper and lower face AUs and AU combinations on the landmark localization rates will be analyzed and discussed.

4.1 Effect of Upper Face AUs on Eye Region Localization Rates

On the average, the results demonstrated that eye region localization was robust in some extent with

Table 2: AU groups for analysis of the effect of upper and lower face AUs on the method performance.

AU groups	AU subgroups
I. Upper face AUs	1, 2, 4, 5, 6, 7, 43&45
II. Lower face AUs	9, 10, 11, 12, 14, 15, 16, 17, 18, 20, 23, 24, 25, 26, 27
III. Upper face AU combinations	1+2, 1+4, 1+5, 1+6, 1+7, 2+4, 2+5,4+5, 4+6, 4+7, 4+45, 6+7
IV. Lower face AU combinations	9+17, 9+23, 9+25, 10+17, 10+20, 10+25 11+20, 11+25, 12+16, 12+20, 12+25, 15+17, 15+24, 16+20, 16+25, 17+23, 17+24, 17+25, 18+23, 20+25, 23+24, 25+26

Table 3: Effect of upper and lower face AUs and AU combinations on the eye region localization rates.

Effect	I. Upper face AUs	II. Lower face AUs	III. Upper face AU combinations	IV. Lower face AU combinations
No or Slight	1, 2, 5	11, 12, 14, 15, 16, 20, 25, 26, 27	1+2, 1+4, 1+5, 1+6, 1+7, 2+4, 2+5, 4+5	10+20, 10+25, 11+20, 11+25, 12+16, 12+20, 12+25, 15+17, 15+24, 20+25, 25+26, 25+27
Medium	4, 6, 43&45	17, 18, 23, 24	-	9+23, 16+20, 16+25, 17+24, 17+25, 18+23
Strong	**7**	**9, 10,**	**4+6, 4+7, 4+45, 6+7**	**9+17, 9+25, 10+17, 17+23, 23+24**

Table 4: Effect of upper and lower face AUs and AU combinations on the nose and mouth localization rates.

Effect	I. Upper face AUs	II. Lower face AUs	III. Upper face AU combinations	IV. Lower face AU combinations
No or Slight	-	-	-	-
Medium	2m	27	(1+2)m, (1+5)m, (2+5)m	15+24, 25+27
Strong	1, 2n, 4, 5, 6, **7**, 43&45	**9**, **10**, 11, 12, 14, 15, 16, 17, 18, 20, 23, 24, 25, 26	(1+2)n, 1+4, (1+5)n, 1+6, 1+7, 2+4, (2+5)n, 4+5, **4+6**, **4+7**, **4+45**, **6+7**	**9+17**, 9+23, **9+25**, **10+17**, 10+20, 10+25, 11+20, 11+25, 12+16, 12+20, 12+25, 15+17, 16+20, 16+25, **17+23**, 17+24, 17+25, 18+23, 20+25, **23+24**, 25+26

Note: Letters n and m indicate different localization results for nose and mouth localization.

respect to facial expressions. Thus, upper face AUs (1, 2 and 5) and AU combinations (1+2, 1+4, 1+5, 1+6, 1+7, 2+4, 2+5, 4+5) which result in raising of eyebrows and widening of eyelids had a slight or no effect on the eye region localization. The degradation in the eye region localization rates was mainly caused by activation of upper face AUs (4, 6, 7, and 43/45) and AU combinations (4+6, 4+7, 4+45, and 6+7) which typically narrow down a space between the eyelids and/or cause the eyebrows to draw down together. These facial behaviours were the main reasons for wrong eye region localization error.

Recently, studies on the feature-based AU recognition, which performance depends on the features used, reported similar results. In (Lien, Kanade, Cohn, and Li, 2000), first-order derivative filters of different orientations (horizontal, vertical, and diagonal) were utilized to detect transient facial features (wrinkles and furrows) for the purpose of AU recognition. They reported AU recognition rate of 86% for AU 1+2, 80% for AU1+4, and 96% for AU4. In (Tian, Kanade, and Cohn, 2002), the authors reported a decrease in performance of the feature-based AU recognition for nearly all the same AUs (AU 4, 5, 6, 7, 41, 43, 45, and 46) which created difficulties in landmark localization in the present study. Among all the upper face AUs, they found AUs 5, 6, 7, 41, and 43 as the most difficult to process with feature-based AU recognition method.

4.2 Effect of Lower Face AUs on Nose and Mouth Localization Rates

The results demonstrated that nose and mouth localization was significantly affected by facial expressions in both upper and lower face. As it was suggested in (Guizatdinova and Surakka, 2005), AUs 9, 10, 11, and 12 were found to cause a poor localization performance of the method.

There are certain changes in the face when the listed AUs are activated. In particular, when AU12 is activated, it pulls the lips back and obliquely upwards. Further, the activation of AUs 9 and 10 lift the centre of the upper lip upwards making the shape of the mouth resemble an upside down curve. AUs 9, 10, 11, and 12 all result in deepening of the nasolabial furrow and pulling it laterally upwards. Although, there are marked differences in the shape of the nasolabial deepening and mouth shaping for these AUs, it can be summed up that these AUs generally make the gap between nose and mouth smaller. These changes in the facial appearance typically caused wrong nose and mouth localization errors.

Especially, lower face AU 9 and AU combinations 4+6, 9+17, 12+20, 12+16 caused strong degradation in nose and mouth localization rates. Similarly, in (Lien, Kanade, Cohn, and Li, 2000), degradation in the feature-based recognition of the lower face AU combinations 12+25 and 9+17 was observed (84% and 77%, respectively). However, regardless of considerable deterioration of nose and mouth localization by the listed AUs, mouth could be found regardless of whether the mouth was open or closed and whether the teeth or tongue were visible or not (Figure 2).

4.3 General Discussion

So far we discussed the effect of upper face AUs on the eye region localization and the effect of lower face AUs on the nose and mouth localization. However, the results also revealed that expressions in the upper face noticeably deteriorated nose and mouth localization and some changes in the lower face affected eye region localization. It is due to the fact that occurring singly or in combinations, AUs may produce strong skin deformations to be in a far neighbourhood from those AUs. In the current database, upper face AUs were usually represented in conjunction with lower face AUs, and their joint activation caused changes in both upper and lower parts of the face. Because of this, the effect of single AU or AU combinations was difficult to bring into the light. The present study investigated only the indirect effect of AUs and AU combinations on the landmark localization.

The overall performance of the method can be improved in several respects. First, the results demonstrated that a majority of the errors was caused by those facial behaviours which resulted in the decrease of space between neighbouring landmarks. Thus, wrong localization errors occurred already on the stage of edge map construction. The reason for that was that a distance between edges extracted from neighbouring landmarks became less than a fixed threshold and edges belonging to different landmarks were erroneously grouped together. To fix this problem, adaptive thresholds are needed for edge grouping. To facilitate landmark localization further, the merged landmarks can be analyzed according to edge density inside the merged regions. The results showed that the regions of merged landmarks have non-uniform edge density. Such regions can be processed subsequently and separated into several regions of strong edge concentration. Second, it is widely accepted that analysis of spatial semantics among neighbouring facial features helps in detecting and inferring missed or occluded facial landmarks. To improve the performance of the method, a constellation of landmark candidates can be analyzed according to face geometry at the stage of orientation matching. As the results showed, eye regions were localized robustly regardless of facial expression. It gives a possibility to use eye region locations and overall face geometry as a guide for localization of other landmarks which were missed (occluded). It can also decrease a false localization rate.

In summary, the method was effective in localization of facial landmarks in neutral images. In this case, the localization rates were higher than 90% for all facial landmarks. In case of expressive faces, the present results specified some of the critical facial behaviours that caused the degradation of the landmark localization rates. We believe that these results can be generalized in some extent to other methods of landmark detection which rely on the low-level edge and intensity information. Further, using only grey level information contained in the image, the method was invariant with respect to different skin colour. The edge orientation model appeared to be effective in noise reduction. Thus the method was able to locate landmarks in images with hair and shoulders. Emphasizing simplicity and low computation cost of the method, we conclude that it can be used in the preliminary localization of regions of facial landmarks for their subsequent processing where coarse landmark localization is following by fine feature detection.

ACKNOWLEDGEMENTS

This work was financially supported by the Finnish Academy (project number 177857), the University of Tampere, and the Tampere Graduate School in Information Science and Engineering. The authors thank the creators of the Cohn-Kanade AU-Coded Facial Expression Database for the permission to reprint the examples of expressive images.

REFERENCES

Ekman, P., Friesen, W., 1978. *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action,* Consulting Psychologists Press, Inc. Palo Alto, California.

Ekman, P., Friesen, W., Hager, J., 2002. *Facial Action Coding System (FACS),* A Human Face. Salt Lake City, Utah.

Gizatdinova, Y., Surakka, V., 2006. Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (1), pp. 135-139.

Guizatdinova, I., Surakka, V., 2005. Detection of Facial Landmarks from Neutral, Happy, and Disgust Facial Images. In *Proceedings of 13th Int. Conf. Central Europe on Computer Graphics, Visualization and Computer Vision*, pp. 55-62.

Hjelmas, E. Low, B., 2001. Face Detection: A Survey. In *Computer Vision and Image Understanding*, 83, pp. 235–274.

Kanade, T., Cohn, J., Tian, Y., 2000. Comprehensive Database for Facial Expression Analysis. In *Proceedings of 4th IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 46-53.

Lien, J., Kanade, T., Cohn, J., Li, C., 2000. Detection, Tracking, and Classification of Action Units in Facial Expression. In *J. Robotics and Autonomous Systems*, 31, pp. 131-146.

Pantic, M., Rothkrantz, J., 2000. Automatic Analysis of Facial Expressions: The State of the Art. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22 (12), pp. 1424–1445.

Tian, Y.-L., Kanade, T., Cohn, J., 2002. Evaluation of Gabor Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity. In *Proceedings of 5th IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 229-234.

Yang, M., Kriegman, D., Ahuaja, N., 2002. Detecting Face in Images: A Survey. In *IEEE Trans. Pattern Analysis and Image Understanding*, 24, pp. 34-58.

APPENDIX A: EDGE DETECTION AND GROUPPING

The grey scale image representation was considered as a two dimensional array $I = \{b_{ij}\}$ of the $X \times Y$ size. Each b_{ij} element of the array represented b intensity of the $\{i, j\}$ image pixel. If there was a colour image, it was first transformed into the grey scale representation by averaging of the three RGB components. This allowed the method to be robust with respect to small illumination variations and skin colour. The high frequencies were removed by convolving the image with a Gaussian filter to eliminate noise and small details (Equation 1).

$$b_{ij}^{(l)} = \sum_{p,q} a_{pq} b_{ij}^{l-1} \ , \ b_{ij}^{(1)} = b_{ij} \tag{1}$$

where a_{pq} is a coefficient of the Gaussian convolution; p and q define the size of a filter, $p,q = -2 \div 2$; $i = 0 \div X - 1$; $j = 0 \div Y - 1$; $l = 1{,}2$ define the level of image resolution.

The smoothed images were further used to detect regions of image which were more likely to contain facial landmarks. The original, high resolution images were used to analyse the candidates for facial landmarks in more detail. In that way, the amount of information that was processed at high resolution level was significantly reduced.

Further, local oriented edges were extracted by convolving the image with a set of ten convolution kernels resulting from differences of two oriented Gaussians (Equations 2-5).

$$G^{-}_{\varphi_k} = \frac{1}{2\pi\sigma^2} e^{-\frac{(p-\sigma\cos\varphi_k)^2+(q-\sigma\sin\varphi_k)^2}{2\sigma^2}} \tag{2}$$

$$G^{+}_{\varphi_k} = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{(p+\sigma\cos\varphi_k)^2+(q+\sigma\sin\varphi_k)^2}{2\sigma^2}} \tag{3}$$

$$G_{\varphi_k} = \frac{1}{Z}(G^{-}_{\varphi_k} - G^{+}_{\varphi_k}) \tag{4}$$

$$Z = \sum (G^{-}_{\varphi_k} - G^{+}_{\varphi_k}) \ , \ G^{-}_{\varphi_k} - G^{+}_{\varphi_k} > 0 \tag{5}$$

where $\sigma = 1.2$ is a root mean square deviation of the Gaussian distribution; φ_k was an angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$; $k = 2 \div 6{,}10 \div 14$; $p,q = -3 \div 3$.

The maximum response of all 10 kernels defined the contrast magnitude of a local edge at its pixel location (Equation 6). The orientation of a local edge was estimated with orientation of a kernel that gave the maximum response.

$$g_{ij\varphi_k} = \sum_{p,q} b^{(l)}_{i-p,j-q} G_{\varphi_k} \tag{6}$$

After the local oriented edges were extracted, they were thresholded, and then grouped into the regions of interest representing candidates for facial landmarks. The threshold for contrast filtering of the extracted edges was defined as an average contrast of the smoothed image. Edge grouping was based on the neighbourhood distances between edge points and was limited by a number of possible neighbours for each edge point. Regions with small number of edge points were removed. The optimal thresholds for edge grouping were determined using a small image set randomly selected from the database.

To get more detailed description of the extracted edge regions, the steps of edge extraction and edge grouping were applied to high resolution image ($l = 1$) within the limits of these regions. In this case, the threshold for contrast filtering was determined as a double average contrast of the high resolution image.

APPENDIX B: EDGE ORIENTATION MATCHING

The procedure of edge orientation matching was applied to verify the existence of a landmark on the image. To do that, the detected regions were matched against the edge orientation model. The orientation model defined a specific distribution of the local oriented edges inside the detected regions.

The following rules defined the edge orientation model: 1) horizontal orientations are represented by the greatest number of the extracted edges; 2) a number of edges corresponding to each of horizontal orientations is more than 50% greater than a number of edges corresponding to any other orientations; and 3) orientations cannot be represented by zero number of edges.

The regions of facial landmarks had the specific distribution of the oriented edges. On the other hand, non-landmark regions like, for example, elements of clothing and hair, usually had an arbitrary distribution of the oriented edges and were discarded by the orientation model.

Publication IV

Gizatdinova Y. and Surakka V. (2007). Automatic Detection of Facial Landmarks from AU-Coded Expressive Facial Images. *Proceedings of International Conference on Image Analysis and Processing (ICIAP'07)*, IEEE Computer Society, 419-424.

Available online at:
http://ieeexplore.ieee.org (requires subscription)

Automatic Detection of Facial Landmarks from AU-coded Expressive Facial Images

Yulia Gizatdinova and Veikko Surakka
Research Group for Emotions, Sociality, and Computing
Tampere Unit for Computer-Human Interaction
Department of Computer Sciences
University of Tampere
{yulia.gizatdinova, veikko.surakka}@cs.uta.fi

Abstract

The present aim was to develop a fully automatic feature-based method for expression-invariant detection of facial landmarks from still facial images. It is a continuation of our earlier work where we found that some certain muscle contractions made a deteriorating effect on the feature-based landmark detection especially in the lower face. Taking into account this crucial facial behavior, we introduced improvements to the method that allowed facial landmarks to be fully automatically detected from expressive images of high complexity. In the method, information on local oriented edges was utilized to compose edge maps of the image at two levels of resolution. The landmark candidates resulted from this step were further verified by edge orientation matching. We used knowledge on face geometry to find the proper spatial arrangement of the candidates. The results obtained demonstrated a high overall performance of the method while testing a wide range of facial displays.

1. Introduction

Human faces constitute a class of objects with rigid structure that does not vary significantly from person to person (i.e. nose is located between eyes and mouth). However, the problem of automatic detection of face and facial features has been challenging computer scientists already for several decades, and still needs further investigation. The difficulty comes from the fact that facial appearance varies noticeably with changes in environmental conditions (e.g. illumination, head pose, orientation, and occlusions), race, gender and facial expressions (e.g. emotional and social signals in the face). To solve the problem, a representation of the face is needed that remains robust with respect to variety of facial appearances. Following this idea, many techniques to face and facial feature detection have been proposed [1], [2].

In expressive facial behavior, muscle contractions produce skin displacements that change drastically the appearance of permanent (e.g. eyes, eyebrows, nose, and mouth) and transient (e.g. wrinkles resulting from expressive and edge-specific face modifications) facial features. Facial expressions result in considerable changes of feature shapes and their locations on the face, presence/absence of teeth, out-of-plan changes (e.g. showing the tongue), and self-occlusions.

In the domain of behavioral science research, the Facial Action Coding System (FACS) [3], [4] is a well known linguistic description of all visibly detectable changes in the facial appearance. The FACS describes visible changes in the face as a result of single and joint muscle contractions in terms of action units (AUs). In other words, FACS represents an expressive image as a result of facial muscle activity without referring to emotional state of a person on the image.

Addressing the problem of expression-invariant facial landmark detection, Gizatdinova and Surakka [5] introduced a feature-based method that made use of local oriented edges extracted in still facial image. For this purpose, a set of multiorientation and multiresolution Gaussian filters was utilized. The detailed description of edge detection and grouping used can be found in Appendix A. Resulting from these stages, the final edge map of the image consisted of regions of connected edges presuming to contain facial landmarks. The existence of a landmark on the image was verified by matching candidates against the orientation model (for more details, see Appendix B).

The method was not fully automatic and required a manual classification of the detected edge regions. Besides that, the method was deteriorated by facial

0-7695-2877-5/07 $25.00 © 2007 IEEE

expressions, especially by those appeared in lower face [6]. The further analysis [7] revealed specific facial behaviors that influenced the performance of the method the most. It was found that incorrect nose and mouth detection was caused mainly by AUs activated during disgust (AU 9 and 10), happiness (AU 12), and some of their combinations with other AUs. Although the listed AUs have different effect on facial appearance, they commonly make the gap between nose and mouth smaller. The neighborhood distances between edges belonging to these landmarks became smaller than a threshold and caused erroneous grouping of nose and mouth into one region. In some cases, AUs 1 and 4 activated during sadness and anger caused eyes or eyebrows to draw up together resulting in incorrect upper face landmark detection.

In the present study, we extended the previous research. Taking into account the described facial behaviors interfering landmark detection, we improved the overall performance of the method. The method now allowed facial landmarks to be fully automatically detected from expressive images of high complexity.

2. Facial landmark detection

The method was improved in several respects. Instead of using an average contrast of the whole image to define thresholds for contrast filtering, we applied local contrast thresholding calculated in every filter neighborhood. This allowed more reliable edge detection. Further, edge grouping was improved as the method failed at this stage due to erroneous connection of edges belonging to different facial landmarks into one region. The top row of Figure 1 shows bounding box that includes merged eye regions on the left and merged nose and mouth on the right. To fix this problem, we applied the procedure of edge projection as follows. If a landmark candidate consisted of two or more regions of edge concentration, edge points were projected to x-axis for upper face landmarks and to y-axis for lower face landmarks. The projections were obtained from calculating the number of edge points along the corresponding (i.e. vertical or horizontal) rows of the final edge map for the given candidate. If the number of edge points was smaller than a threshold, edge points were eliminated (Figure 1). After each edge elimination step, if the region still was not separated the threshold was increased by 5 edge points. The initial threshold equaled a minimum number of edges in the column (row) of the given candidate.

After the procedure of edge projection, the orientation portraits (i.e. the distribution of local oriented edges) of the received edge regions were matched against the orientation model. In this study, we allowed landmark candidates to have some deviations from the orientation model. It means that an orientation portrait of the candidate could slightly differ from the model, for example, it could have some orientations represented by zero number of edges. In further analysis, these edge regions were also considered in composing face-like constellations of the detected landmark candidates if there were missing landmarks. Figure 2,a shows the final edge map of the image with landmark candidates and discarded edge regions.

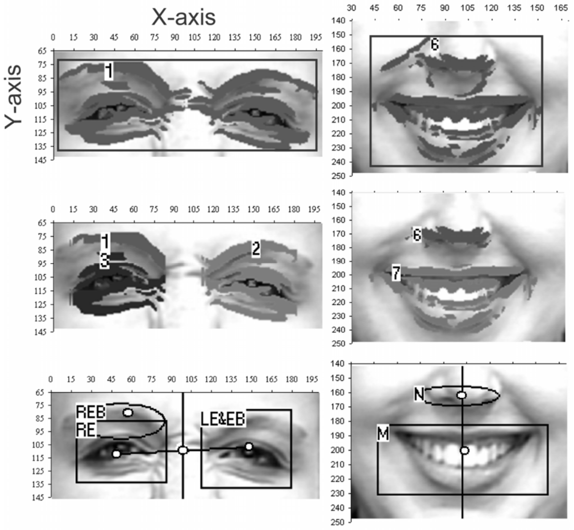


Figure 1. Landmarks grouped into one region (top), landmarks separated by edge projection (middle), and final detection result (bottom). Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression database [8]. Reprinted with permission.

The final improvement of the method was the automatic classification of the detected landmark candidates. We formed constellations from a set of detected candidates and determined which constellations were the most face-like. The face model we used is shown in Figure 2,b. Due to side-by-side location of upper face landmarks, they guided the entire process of landmark classification, also those landmarks which were discarded by the orientation model. The search started with finding horizontal candidate pairs with approximately equal number of edge points and labeling them as eyes and eyebrows. If only one horizontal pair was found, it was labeled as eye region candidate (i.e. eye and eyebrow were detected as one region). The method then searched for

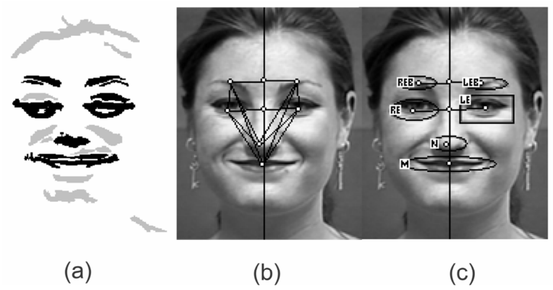


Figure 2. (a) Final edge map with landmark candidates (black) and discarded edge regions (grey), (b) face geometry model, and (c) final detection result. Image is courtesy of Cohn- Kanade AU-Coded Facial Expression database [8]. Reprinted with permission.

eyebrows above and eyes below the found pair location and if found any, relabeled the found candidates as eye and eyebrow, respectively. If there was not any pair found, it was assumed that eye regions were grouped together in one region and edge x-projection was applied to the candidate with maximum number of edges. The search for lower face landmarks was performed from top-to-bottom along the line of vertical symmetry that was drawn through the point that lied in the middle of the line connecting eye regions. If only one lower face candidate was found, the method assumed that nose and mouth were combined together and edge y-projection was applied to separate these landmarks. Although the method was allowed to miss landmarks, however, for efficient landmark detection at least one horizontal pair had to be found. As a measure of distances in the face model we utilized the dynamic parameter D calculated as a distance between mass centers of the eye region pair. Using this measure, the spatial constraints between locations of the rest of the candidates were verified. For example, nose is located between eyes and mouth not lower than one D from the middle point of the line connecting eye regions. At the same time, by utilizing geometrical relationships among the candidates, we verified the upper face landmarks. After the face-like constellation of landmarks was found, the location of the face in the image was also known.

3. Database

The Cohn-Kanade AU-Coded Facial Expression Database [8] consists of image sequences taken from 97 subjects (65% female) of different skin color (81% Caucasian, 13% African-American, and 6% Asian or Latino) and ages varying from 18 to 30 years. There were no images with facial hair or eye-glasses. Each image sequence starts with neutral frame and ends up with an expressive frame labeled in terms of AUs. AUs occur alone or in combinations and are coded as numbers. The level of expression intensity can vary for images of different subjects and is coded as small letters. Capital letters L and R define left- and right-side expressions.

From the database we selected 468 neutral and 468 expressive images corresponding to the first and the last frames of the sequence. From this data we composed two datasets – "face only" dataset of cropped images including only facial region, and "face & hair" dataset of cropped images including both face and hair. "Face only" dataset served as a "baseline" to which we compared the robustness of the method with respect to such destructors as hair, decoration, and elements of clothe. All images were preset to the size of approximately 200-250 pixel arrays with 8-bit precision for grey scale values. No face alignment was performed.

4. Results

The following facial landmarks were chosen to be detected: - right eye (RE), right eyebrow (REB), left eye (LE), left eyebrow (LEB), right eye and eyebrow (RE&EB), left eye and eyebrow (LE&EB), lower nose (N), and mouth (M). Figure 3 shows the final results of the landmark detection in both datasets. As figure shows, the size of the bounding box that contained a landmark was dynamic and varied according to the size of the detected edge region. The landmarks with orientation portraits slightly different from the orientation model were represented as ovals, and landmarks corresponded to the orientation model – as rectangles.

The final results were classified into one of the following classes: correct detection, wrong detection, and false detection. A correct detection was considered if the bounding box overlapped approximately at least 50% of the visible landmark, and edge region enclosed the area surrounding landmark less than the actual size of the landmark. In detecting eye regions, eyebrow together with a corresponding eye were localized as one region, or alternatively, eye and eyebrow were localized separately. If eyebrow was detected as a separate region, it was obligatory that a corresponding eye was also found. A wrong detection was considered if the bounding box covered several facial landmarks, excluding the case of eyes and eyebrows localized as one region. A false localization was considered when bounding box did not satisfy any of the two previous conditions. We defined the rate of the landmark detection as a ratio between a total number of

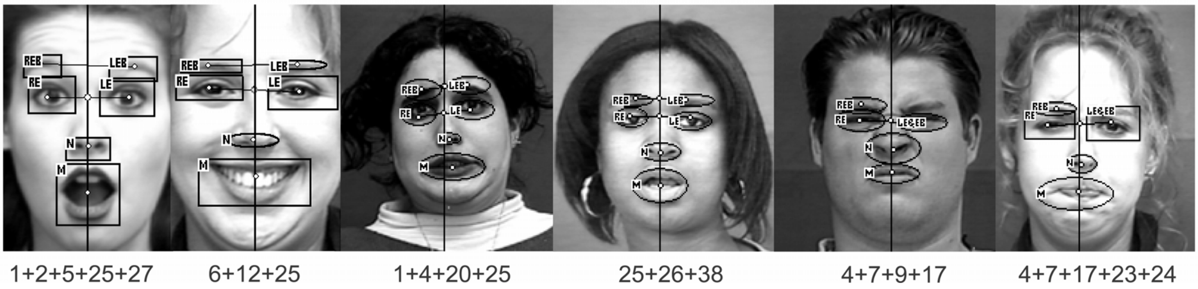


Figure 3. Examples of correctly localized facial landmarks in "face only" and "face & hair" datasets. Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression database [8]. Reprinted with permission.

landmarks correctly localized and the total number of images used in testing (as there was one face per image). A false positive was then defined as a number of noise regions (wrinkles, eyebrow localized without a corresponding eye, elements of face, ears, clothing and hair) which were misclassified as a facial landmark.

As it is seen from Table 1 and Table 2, there was no significant difference in the performance of the method on two datasets. Further, the facial landmarks were detected with nearly equal detection rates in both neutral and expressive images. Thus, the method achieved the average detection rates of 97.5% and 94% for neutral and expressive "face only" images, correspondently. The rates for "face & hair" dataset were 91.5% for neutral images and 90% for expressive images. A decrease in detection rates for lower face landmarks was observed; on the whole, however, the overall performance of the method was high.

We noticed that detection of lower face landmarks produced more errors than detection of upper face landmarks. For example, in some cases the method misclassified a chin as a mouth, (in the tables, the biggest number of false positives corresponding to mouth detection reflects this fact). Wrong detections were observed mostly in detecting lower face landmarks. Thus, nose and mouth were detected as one region in 16 expressive images of "face only" dataset and in 18 expressive images of "face & hair" dataset. As it was expected, it occurred mainly due to the effect of lower face AUs 9, 10, and 12 occurring alone or in combinations with other AUs.

Table 1. Landmark detection rates (%) and false positives (FP) for "face only" dataset

Image	Right eye region	Left eye region	Nose	Mouth
Neutral	98 1 FP	99 1 FP	98 0 FP	95 12 FP
Expressive	97 6 FP	98 2 FP	92 1 FP	90 12 FP

Table 2. Landmark detection rates (%) and false positives (FP) for "face & hair" dataset

Image	Right eye region	Left eye region	Nose	Mouth
Neutral	95 2 FP	96 3 FP	89 1 FP	86 19 FP
Expressive	93 5 FP	94 6 FP	90 2 FP	81 25 FP

The eye region detection was high for all types of images showing expressions in upper and lower face, see Table 3 and Table 4, (note that AUs presented might occur singly or in conjunction with other AUs which are not represented in the tables). On the whole, the detection of lower face landmarks was more affected by AUs than the detection of eye regions. Lower face AUs 9, 10, 12, and AU combinations 9+25, 10+17, 10+20, 10+25, 12+16, 12+25, and 16+25 lowered down the nose and mouth detection average rates up to the range of 71-83%. Upper face AUs 5, 6, 7, and AU combination 6+7 also degraded the lower face landmark detection. These upper face AUs are usually activated during the lower face expression of anger when AUs 9 and 10 typically are also activated.

Table 3. Rates (%) of landmark detection in images showing upper face single AUs

AUs	Right eye region	Left eye region	Nose	Mouth
1	92	96	90	83
2	90	96	94	88
4	94	95	91	86
5	89	93	91	73
6	95	93	86	77
7	93	96	86	73
43/45	95	98	88	88

Table 4. Rates (%) of landmark detection in images showing lower face single AUs

AUs	Right eye region	Left eye region	Nose	Mouth
9	93	94	87	81
10	94	89	82	78
11	98	98	87	87
12	96	94	79	71
14	84	100	89	89
15	96	96	95	91
16	97	100	94	90
17	94	95	92	91
20	97	94	94	89
23	95	93	93	92
24	94	93	94	94
25	95	96	91	92
26	94	98	97	94
27	92	97	96	88

5. Discussion

A fully automatic method was designed for facial landmark detection in the expressive images of high complexity. The complexity of the expression was presented by closed/semi-closed eyes, variety of mouth appearances including open and tight mouth, visible teeth and tongue. The local oriented edges served as basic features for expression-invariant representation of facial landmarks. The results confirmed that in the majority of expressive images the landmark orientation portraits had the same structure as predefined by the landmark orientation model. The face geometry model further improved the overall performance of the method. Besides robustness to facial expressions, the method demonstrated robustness to skin color and noise like hair, ear-rings, and elements of clothe.

Comparing present results with previous ones, a significant improvement was achieved for detection of lower face landmarks, especially, in images showing AUs 9, 10, and 12. The landmark detection rates were comparable or superior to those presented in [9]-[11] while testing a wider range of facial displays.

Emphasizing simplicity of the method developed, we conclude that it can be used in preliminary localization of regions of facial landmarks for their subsequent processing where coarse landmark localization is followed by fine feature detection (e.g. local features like eye and mouth corners).

6. Acknowledgement

This work was financially supported by the Finnish Academy (project number 177857), the University of Tampere, and the Tampere Graduate School in Information Science and Engineering. The authors thank the creators of the Cohn-Kanade AU-Coded Facial Expression Database for the permission to reprint examples of the expressive images.

7. References

[1] E. Hjelmas, B. Low, "Face Detection: A survey", *Computer Vision and Image Understanding*, 83, 2001, pp. 235–274.

[2] M. Yang, D. Kriegman, and N. Ahuaja, "Detecting Face in Images: A Survey", *IEEE Transactions on Pattern Analysis and Image Understanding*, 24, 2002, pp. 34-58.

[3] Ekman, P., W. Friesen, *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action*, Consulting Psychologists Press, Inc., Palo Alto, California, 1978.

[4] Ekman, P., W. Friesen, and J. Hager, *Facial Action Coding System (FACS)*, A Human Face, Salt Lake City, Utah, 2002.

[5] Y. Gizatdinova, V. Surakka, "Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 1, 2006, pp. 135-139.

[6] I. Guizatdinova, V. Surakka, "Detection of Facial Landmarks from Neutral, Happy, and Disgust Facial Images", *Proceedings of the 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, Plzen, Czech Republic, 2005, pp. 55-62.

[7] Y. Gizatdinova, V. Surakka, "Edge Orientation Matching for Facial Landmark Localization in Images Showing Expressions in Upper and Lower face", submitted to *Computer Vision and Image Understanding.*

[8] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis", *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp. 46-53.

[9] P. Campadelli, R. Lanzarotti, G. Lipori, and E. Salvi, "Face and Facial Feature Localization", *Proceedings of the 13th International Conference on Image Analysis and Processing (ICIAP)*, Gagliari, Italy, 2005, 3617, pp. 1002-1009.

[10] D. Shaposhnikov, A. Golovan, L. Podladchikova, N. Shevtsova, X. Gao, V. Gusakova, and Y. Gizatdinova, "Application of the Behavioral Model of Vision for Invariant Recognition of Facial and Traffic Sign Images" (Применение поведенческой модели зрения для инвариантного распознавания лиц и дорожных знаков), *Journal of Neurocomputers: Design and Application*, 7(8), 2002, pp. 21-33.

[11] D. Cristinacce, T. Cootes, “Facial Feature Detection Using AdaBoost with Shape Constraints”, *Proceedings of the 14th British Machine Vision Conference (BMVC)*, Norwich, England, 2003, pp. 231-240.

8. Appendix A: Edge detection

The grey scale image representation was considered as a two dimensional array $I = \{b_{ij}\}$ of the $X \times Y$ size. Each b_{ij} element of the array represented b brightness of the $\{i, j\}$ image pixel. If there was a color image, it was first transformed into the grey scale representation by averaging three RGB components. This allowed the method to be robust with respect to small illumination variations and skin color. To smooth a grey level image the recursive Gaussian transformation was used.

$$b_{ij}^{(l)} = \sum_{p,q} a_{pq} b_{ij}^{l-1}\ ,\ b_{ij}^{(1)} = b_{ij}\ , \tag{1}$$

where a_{pq} was a coefficient of the Gaussian convolution; p and q defined the size of a filter, $p, q = -2 \div 2$; $i = 0 \div X - 1$; $j = 0 \div Y - 1$; l defined the level of image resolution. The smoothed low resolution image (*l=2*) was used to find all possible landmark candidates, and the original high resolution image (*l=1*) was used to analyse landmark candidates in detail.

Then the smoothed image was convolved with a set of ten-orientation Gaussian filters with shifted centres.

$$G_{\varphi_k}^{-} = \frac{1}{2\pi\sigma^2} e^{-\frac{(p-\sigma\cos\varphi_k)^2+(q-\sigma\sin\varphi_k)^2}{2\sigma^2}}\ , \tag{2}$$

$$G_{\varphi_k}^{+} = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{(p+\sigma\cos\varphi_k)^2+(q+\sigma\sin\varphi_k)^2}{2\sigma^2}}\ , \tag{3}$$

$$G_{\varphi_k} = \frac{1}{Z}(G_{\varphi_k}^{-} - G_{\varphi_k}^{+})\ , \tag{4}$$

$$Z = \sum (G_{\varphi_k}^{-} - G_{\varphi_k}^{+})\ ,\ \ G_{\varphi_k}^{-} - G_{\varphi_k}^{+} > 0\ , \tag{5}$$

where σ was a root mean square deviation of the Gaussian distribution; φ_k was an angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$; $k = 2 \div 6, 10 \div 14$; $p, q = -3 \div 3$.

The maximum response of all 10 kernels defined the contrast magnitude of a local edge at its pixel location. The orientation of a local edge was estimated with orientation of a kernel that gave the maximum response.

$$g_{ij\varphi_k} = \sum_{p,q} b_{i-p,\,j-q}^{(l)} G_{\varphi_k}\ , \tag{6}$$

The threshold for contrast filtering of the extracted edges was determined as an average contrast of the whole smoothed image. Edge grouping was based on neighborhood distances between edge points and limited by a radius consisting of possible neighbors for each edge point. Regions with small number of edge points were removed. The optimal thresholds for edge grouping were determined using small image set taken from the database. To get more detailed description of the extracted edge regions, edge detection and grouping were applied to high resolution image within the limits of these regions. In this case, the threshold for contrast filtering was determined as a double average contrast of the high resolution image.

9. Appendix B: Edge orientation matching

The procedure of orientation matching was applied to verify the existence of facial landmarks on the image. To do that, the detected regions were matched against the orientation model that was a specific distribution of the local oriented edges with two horizontal dominants (for example, see Figure 4). The following rules define the distribution of the orientation model: 1) horizontal orientations are represented by the greatest number of the extracted edges; 2) a number of edges corresponding to each of horizontal orientations is more than 50% greater than a number of edges corresponding to any other orientations; and 3) orientations cannot be represented by zero number of edges. Noise regions like, for example, elements of cloth and hair usually have an arbitrary distribution of the oriented edges and were discarded by the model.

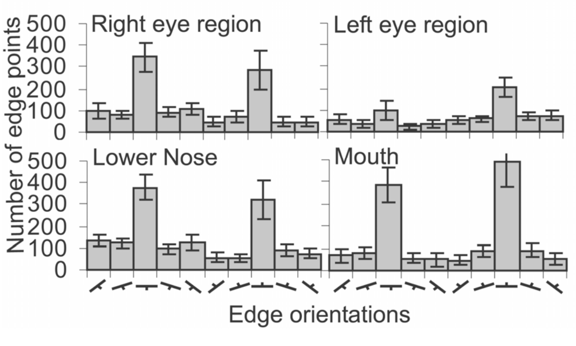
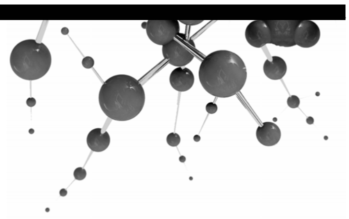


Figure 4. Examples of landmark orientation portraits averaged over “face only” datasets. The error bars show plus/minus one standard deviation from the mean values.



Publication V

Gizatdinova Y. and Surakka V. (2008). Automatic Detection of Facial Landmarks from Expressive Images of High Complexity. *Technical report D-2008-9*, Department of Computer Sciences, University of Tampere, 1-23.

Available online at:
http://www.cs.uta.fi/reports/sarjad.html

Yulia Gizatdinova and Veikko Surakka

Automatic localization of facial landmarks from expressive images of high complexity

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF TAMPERE

D-2008-9

TAMPERE 2008

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER SCIENCES
SERIES OF PUBLICATIONS D – NET PUBLICATIONS
D-2008-9, SEPTEMBER 2008

Yulia Gizatdinova and Veikko Surakka

Automatic localization of facial landmarks from expressive images of high complexity

DEPARTMENT OF COMPUTER SCIENCES
FIN-33014 UNIVERSITY OF TAMPERE

ISBN 978-951-44-7498-9
ISSN 1795-4274

Automatic Localization of Facial Landmarks from Expressive Images of High Complexity

Yulia Gizatdinova and Veikko Surakka

Abstract

The aim of this study was to develop a fully automatic feature-based method for expression-invariant localization of facial landmarks from static facial images. It was a continuation of our earlier work in which we found that lower face expressions deteriorated the feature-based localization of facial landmarks the most. Taking into account the found crucial facial behaviours, the method was improved so that it allowed facial landmarks to be fully automatically located from expressive images of high complexity. Information on local oriented edges was utilized to compose edge maps of the image at several levels of resolution. Landmark candidates which resulted from this step were further verified by matching them against the orientation model. The present novelty came from two last steps of the method which were edge projection that was used to enhance a search for landmark candidates, and structural correction of the found candidates based on a face geometry model. The method was tested on three facial expression databases which represented a wide range of facial appearances in terms of prototypical facial displays and individual facial muscle contractions. To evaluate the localization results, a new performance evaluation measure was introduced that represented a localization result as a rectangular box instead of a conventional point representation. The evaluation of the method by the proposed rectangular and conventional point evaluation measures demonstrated a high overall performance of the method in localization of facial landmarks from images showing complex expressions in upper and lower face.

Keywords - Image processing, computer vision, edge detection, landmark localization, action unit (AU), facial expression.

I. INTRODUCTION

Facial expressions are emotional, social, and otherwise meaningful cognitive and physiological signals in the face. Facial expressions result from contractions and relaxations of different facial muscles. These non-rigid facial movements result in considerable changes of facial landmark shapes and their location on the face, visibility of teeth, out-of-plan changes (showing the tongue), and self-occlusions (close eyes and bitted lips). In the domain of behavioural science research, facial expressions can be viewed according to two main approaches. The emotion-based approach proposed by Ekman [1] considers facial expressions as primary channels of conveying emotional information. Ekman defined six prototypical emotions and six corresponding prototypical facial expressions which are happiness, sadness, fear, anger, disgust, and surprise. In the contrast to emotion-based approach, the Facial Action Coding System (FACS) [2,3] proposes a descriptive approach to facial expression analysis. The FACS is an anatomically based linguistic description of all visibly detectable changes in the face. The FACS describes visible changes in the face as a result of single and conjoint muscle activations in terms of action units (AUs). In other words, FACS decomposes an expressive face into AUs and represents an expression as a result of facial muscle activity fully objectively, without referring to emotional, social, or cognitive state of the person in the image. However, some specific combinations of AUs can represent prototypical facial expressions of emotional.

It has been shown [1] that structural changes in the regions of prominent facial landmarks like, for example, eyebrows, eyes, nose, and mouth are important and in many cases sufficient for facial expression classification. In the automatic facial expression analysis, a manual preprocessing is typically needed to select a set of characteristic points as, for example, eye centres and mouth corners, in static images or initial frame of the video sequence. These characteristic points are further used to track changes in the face or to align an input image with a face model. Currently, there is a need for a system that can automatically detect facial landmarks in the image prior to the following steps of the automatic facial expression analysis.

The problem of automatic facial landmark detection has been generally addressed by modelling local texture information around landmarks and modelling shape information on spatial arrangement of the detected landmark candidates [4,5,6]. In practice, this process consists of selecting a feature representation of facial landmarks and designing a feature detector. Different features can be detected from the image, for example, edges, colours, points, lines, and contours. These features provide a meaningful and measurable description of the face as they represent specific visual patterns which can be used to identify corresponding structures between images. The main challenge is to find feature representations of facial landmarks which uniquely characterize a face and remain robust with respect to changes in facial appearance due to changes in the environmental conditions (illumination, pose, orientation, etc.), gender, race, and facial expressions. Facial landmark localization is a subtask of a more general detection problem and refers to finding true locations of facial landmarks, given that a face is shown in the image.

To detect facial landmarks, Burl, Leung and Perona [7] modelled a texture around the landmarks by employing a set of multi-scale and multi-orientation Gaussian derivative filters. The most face-like constellation of the found candidate locations was further captured by a statistical shape model. In Wiskott et al. [8], the locations of characteristic points around facial landmarks were first found using Gabor jet detectors. Further, the distribution of facial points was modelled by a graph structure. Feris et al. [9] proposed a two-level hierarchical landmark search using Gabor wavelet networks. In this method, the first level network represented the entire face and determined affine transformation used for a rough approximation of the landmark locations. The second level networks represented separate landmarks and were used to verify the precise landmark locations. Similarly, Cristinacce and Cootes [10] extended a well known face detector introduced by Viola and Jones [11] for the task of detecting individual facial landmarks. The local boosted classifiers were used to detect facial landmarks and statistical models of the landmark configurations were utilized to select the most suitable candidates.

Addressing the problem of facial landmark localization, Gizatdinova and Surakka [12] introduced a feature-based method in which the information on local oriented edges was utilized to compose edge maps of the image at several levels of resolution. Landmark candidates which resulted from this step were further verified by matching them against the orientation model. The method was not fully automatic and required a manual classification of the located edge regions. Besides that, the landmark localization was significantly deteriorated by the lower face expressions [13]. The further analysis [14] revealed specific facial behaviours which influenced the performance of the method the most. It was found that incorrect nose and mouth localization was caused mainly by the lower face AU 12 (lip corner puller) activated during happiness, AU 9 (nose wrinkler) and AU 10 (upper lip raiser) both activated during disgust, and AU 11 (nasolabial furrow deepener) that is usually activated in conjunction with all mentioned AUs.

Taking into account the described facial behaviours which deteriorated the landmark localization, we made a number of improvements to the method which allowed facial landmarks to be fully automatically located from expressive images of high complexity. The preliminary results [15] of the method testing on the AU-coded facial expression database showed a significant improvement of the method performance for images showing lower face AUs 9, 10, 11, and 12. In the present study, we continued improving the method and presented new results of the method testing on a wider range of facial displays classified in terms of prototypical facial expressions, single AUs, and AU combinations. A new performance evaluation measure was introduced in order to evaluate the results of landmark localization. The proposed evaluation measure represented a localization result as a rectangular box instead of a conventional point representation.

II. DATABASES

The first database we used was the Cohn-Kanade AU-Coded Facial Expression (Cohn-Kanade) database [16] - one of the most comprehensive collections of expressive images available. The database consisted of image sequences taken from 97 subjects (65% female) of different skin colour (81% Caucasian, 13% African-American, and 6% Asian or Latino) and ages varying from 18 to 30 years. Each image sequence started with a neutral frame and ended up with an expressive frame labelled in terms of AUs. AUs occurred alone or in combinations and were coded as numbers. The AU descriptors taken from the FACS manual [2,3] were as follows. Upper face AUs: 1-inner brow raiser, 2-outer brow raiser, 4-brow lowerer, 5-upper lid raiser, 6-cheek raiser and lid compressor, 7-lid tightener, 43-eye closure, 45–blink, and 46-wink. Lower face AUs: 9-nose wrinkler, 10-upper lip raiser, 11-nasolabial furrow deepener, 12-lip corner puller, 13–sharp lip puller, 14-dimpler, 15–lip corner depressor, 16–lower lip depressor, 17-chin raiser, 18-lip pucker, 20–lip stretcher, 22–lip funneler, 23–lip tightener, 24–lip presser, 25–lips part, 26–jaw drop, 27–mouth stretch, and 28–lips suck. Capital letters R and L in front of the numerical code indicated right and left face AUs. Small letters after the numerical code represented an intensity level of the expression. For each subject, we selected one neutral face and several expressive faces of the highest intensity which corresponded to the latest frames in each expressive sequence. In sum, a total of 97 neutral and 486 expressive images were selected. From this original data we composed datasets of cropped images with face, hair, and sometimes shoulders included (the background and the image indexes were cut out).

The images of the rest of the two databases were labelled in terms of neutral and six prototypical facial expressions which were happiness, sadness, fear, anger, disgust, and surprise. The Pictures of Facial Affect (POFA) database [17] consisted of 14 neutral and 96 expressive images of 14 Caucasian individuals (57% female). On average, there were 16 images per facial expression. The Japanese Female Facial Expression (JAFFE) database [18] consisted of 30 neutral and 176 expressive images of 10 Japanese females. There were about 30 images per facial expression in average. In POFA and JAFFE databases, a particular expression could vary in its intensity or facial configuration.

All images were preset to approximately 200 by 300 pixel arrays. No face alignment was performed. The potential impact of illumination change, facial hair or eye-glasses was controlled to some extent in all the databases and therefore ignored. The received datasets were used to examine the robustness of the method with respect to facial activity labelled in terms of single AUs, AU combinations, and prototypical facial expressions. The robustness of the method to such destructors as hair, decoration, and elements of clothing was also studied.

III. FACIAL LANDMARK LOCALIZATION

The feature-based method of facial landmark localization consisted of several stages which are illustrated in Figure 1. The image was considered as a two dimensional array $I = \{b_{ij}\}$ of the $X \times Y$ size.

Each b_{ij} element of the array represented b brightness of the $\{i, j\}$ image pixel. On the preprocessing stage, the image was smoothed by the recursive Gaussian transformation (Equation 1) to remove noise and small details from the image. On the following stages of the method, the smoothed low resolution image was used to find all possible landmark candidates, and the original high resolution image was used to analyze the landmark candidates in detail.

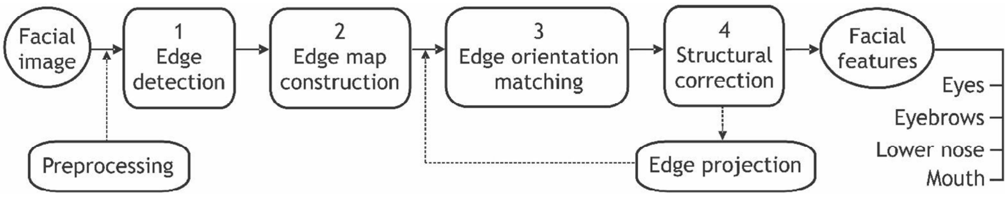


Figure 1. Block structure diagram of the facial landmark localization.

$$b_{ij}^{(l)} = \sum_{p,q} a_{pq} b_{ij}^{l-1}\,, \; b_{ij}^{(1)} = b_{ij}\,. \qquad (1)$$

where a_{pq} is a coefficient of the Gaussian convolution; p and q define a size of the smoothing filter, $p,q = -2 \div 2$; $i = 0 \div X-1$; $j = 0 \div Y-1$; l define a level of resolution ($l = 2$).

If there was a colour image, it was first transformed into the grey scale representation by averaging three RGB components (Equation 2). This allowed the method to be robust with respect to small illumination variations and different skin colour.

$$b_{ij} = 0.299 \cdot R_{ij} + 0.587 \cdot G_{ij} + 0.114 \cdot B_{ij}\,. \qquad (2)$$

On the stage of edge detection, the smoothed low resolution image was filtered with a set of ten-orientation Gaussian filters (Equations 3-6) to extract local oriented edges.

$$G_{\varphi_k} = \frac{1}{Z}(G_{\varphi_k}^{-} - G_{\varphi_k}^{+}), \qquad (3)$$

$$G_{\varphi_k}^{-} = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(p-\sigma\cos\varphi_k)^2 + (q-\sigma\sin\varphi_k)^2}{2\sigma^2}\right), \qquad (4)$$

$$G_{\varphi_k}^{+} = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(p+\sigma\cos\varphi_k)^2 + (q+\sigma\sin\varphi_k)^2}{2\sigma^2}\right), \qquad (5)$$

$$Z = \sum(G^{-}_{\varphi_k} - G^{+}_{\varphi_k}),\ G^{-}_{\varphi_k} - G^{+}_{\varphi_k} > 0. \qquad (6)$$

where σ is a root mean square deviation of the Gaussian distribution; φ_k is an angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$; $k = 2 \div 6,\ 10 \div 14$; $p,q = -3 \div 3$; $i = 0 \div X-1$; $j = 0 \div Y-1$.

The maximum response of all ten kernels (Equation 7) defined a contrast magnitude of the local edge at its pixel location. The orientation of the local edge was estimated by the orientation of the kernel that gave the maximum response.

$$g_{ij\varphi_k} = \sum_{p,q} b^{(l)}_{i-p,j-q} G_{\varphi_k}. \qquad (7)$$

On the stage of edge map construction, the extracted edge points were thresholded according to their contrast. The average contrast of the whole smoothed low resolution image was used to define a threshold for contrast filtering. Edge grouping was based on the neighbourhood distance (D_n) between edge points and limited by a minimum number of edge points in the region ($N_{\min}$). Thus, edge points were grouped into one region if the distance between them was less than D_n pixels and number of edge points inside the region was bigger than $N_{\min}$. Regions with small number of edge points were removed. This way, the final edge map of the image consisted of regions of connected edge points presuming to contain facial landmarks. The optimal thresholds for edge grouping were determined experimentally and summarized in Table I. To get more detailed description of the extracted edge regions, edge detection and edge grouping were applied to high resolution image ($l = 1$) within the limits of the found edge regions. In this case, the threshold for contrast filtering was determined as a double average contrast of the high resolution image.

TABLE I: SUMMARY OF THE THRESHOLDS FOR EDGE MAP CONSTRUCTION

Datasets	Neighborhood distance for edge grouping, D_n	Minimum number of edge points in the region, $N_{\min}$
Cohn-Kanade	1 pixel	100 pixels
POFA	3 pixels	150 pixels
JAFFE	2 pixels	160 pixels

On the stage of edge orientation matching, the existence of facial landmarks in the image was verified. To do that, a distribution of the local oriented edges inside the located regions, so called orientation portraits, was matched against the orientation model. The model specified a characteristic distribution of the local oriented edges with maximums corresponded to two horizontal orientations (dark-to-light and light-to-dark horizontal edges). Unlike facial landmarks, noisy regions as, for example, elements of clothing and hair usually had an arbitrary distribution of the oriented edges and were discarded by the orientation model. Figures 2*b-d* shows the stages 1-3 of the method and were described in more detail in [12].

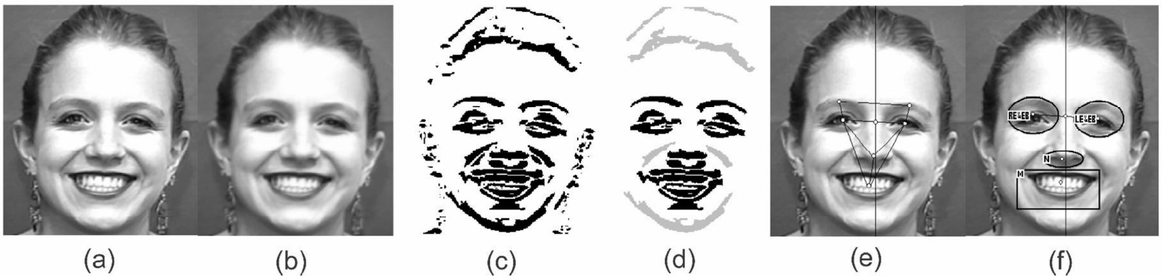


Figure 2. Facial landmark localization: (a) image of happiness; (b) smoothed image; (c) extracted local oriented edges; (d) edges grouped into regions representing landmark candidates (black) and noisy regions discarded by the edge orientation model (grey); (e) face geometry model; and (f) final localization result with "primary" landmarks (rectangles) and "secondary" landmarks (ovals). Image is a courtesy of the Cohn-Kanade database. Reprinted with permission.

A number of improvements were made to the earlier version of the method. First, instead of using a double average contrast of the whole high resolution image to define a threshold for contrast filtering of the located edge regions, we applied local contrast thresholding calculated in every filter neighbourhood. Second, the stage of edge map construction was improved as the method failed at this stage due to erroneous connection of edges belonging to different facial landmarks into one region. The one reason for that was a specific facial behaviour typically caused by AUs 9 (nose wrinkler), 10 (upper lip raiser), 11 (nasolabial furrow deepener), and 12 (lip corner puller). All the listed AUs result in deepening and pulling of the nasolabial furrow laterally up and raising of the upper lip up. Although there are marked differences in the shape of the nasolabial deepening and mouth shaping for these AUs, they all make the gap between nose and mouth smaller. In addition, nose as the most prominent feature in the face, often produces shadows and, thus, creates additional contrast in facial areas located below the nose. These changes in the lower face made the neighbourhood distances between edges extracted from nose and mouth smaller than a fixed threshold and caused a merging of nose and mouth into one region. Another reason for landmarks to be merged were AUs which are activated during anger, disgust, and sadness.

AU 4 (brow lowerer) pulls the eyebrows down and closer together producing vertical wrinkles between them. Further, AU 6 (cheek raiser) and AU 7 (lid tightener) often activated together with AU 4 also create additional contrast around eye regions. AU 9 (nose wrinkler) causes wrinkles to appear along the sides of the nose and across the root of the nose. These facial behaviours resulted in extracting of the noisy edges between the eyes and from the nose bridge region and, in some cases, caused a merging of eyes or eyebrows into one region.

To separate the neighbouring candidates merged into one region we proposed a simple but effective technique of x/y-edge projection. The schematic interpretation of the proposed technique is illustrated in Figure 3. It shows the merged facial landmarks and their separation by x/y-edge projection. If a landmark candidate consisted of two or more regions of edge concentration, edge points were projected to x-axis for upper face landmarks and to y-axis for lower face landmarks. The projections were obtained from calculating a number of edge points along the corresponding columns or rows of the edge map for the upper or lower face landmark candidate, respectively. If the number of projected edge points was smaller than a threshold (bold dashed line in the figures), edge points were eliminated. After each edge elimination step, if the region still was not separated the threshold was increased by 5 edge points. The initial threshold equalled a minimum number of edges in the column or row of the given candidate.

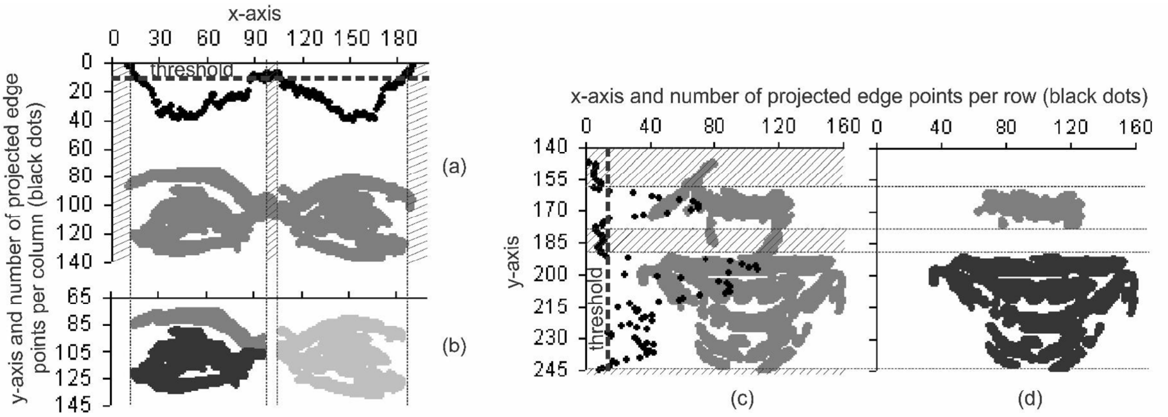


Figure 3. Edge maps of facial landmarks: (a) eye regions wrongly detected as one region; (b) eye regions separated by edge projection; (c) nose and mouth wrongly detected as one region; and (d) nose and mouth separated by edge projection. Black dots represent a number of projected edge points per columns or rows of the upper or lower face landmark candidates, respectively. Areas marked with upward diagonal lines show regions of the edge map where edges were eliminated.

Some changes were also made on the stage of edge orientation matching. The orientation portraits of the landmark candidates were allowed to have slight deviations from the orientation model. In particular, the orientation portrait of the candidate could have not strongly pronounced horizontal dominants in the edge orientation distribution and could have some orientations represented by zero number of edges. If orientation portrait of the candidate corresponded to the model, the candidate was labelled as “primary” candidate. On the contrary, if orientation portrait had slight differences from the orientation model, the corresponded candidate was labelled as “secondary” candidate. In further analysis, “secondary” candidates were also considered in the composition of face-like constellations of the candidates if there were missing landmarks.

Finally, the algorithm of fully automatic classification of the located candidates was applied on the stage of structural correction (see Figure 1). Upper face landmarks to be located were defined as right eye (RE), left eye (LE), right eyebrow (REB), left eyebrow (LEB), right eye region (RER), and left eye region (LER). In two latter cases, eyes and eyebrows were considered as one landmark. Lower face landmarks were represented by lower nose (N) and mouth (M). Due to side-by-side location of the upper face landmarks, their locations guided the entire process of the candidate classification, also those candidates which were discarded by the orientation model. After the face-like constellation of the candidates was found, the location of the face in the image was also known.

In order to find a proper spatial arrangement of the located candidates, the proposed algorithm applied a set of verification rules which were based on the face geometry model (see Figure 2,*e*). The knowledge on face geometry was taken from the anthropological study by Farkas [19]. This thorough study examined thousands of Caucasians, Chinese, and African-American subjects in order to determine characteristic measures and proportion indexes of the human face. The average distance between upper face landmarks (both eyes and eyebrows) *d(RER,LER)* and eyebrow-eye distance appeared to be useful facial measures for the purpose of structural correction of the located candidates. It has been demonstrated that these facial measures can slightly vary between subjects of deferent genders and races. Therefore we defined these measures as intervals between minimum and maximum values for the given database. The defined intervals were [35,85] and [10,35] for Cohn-Kanade, [60,90] and [10,40] for POFA, and [55,80] and [10,30] for JAFFE. The algorithm had several steps which are described in Figure 4.

1. The search started with finding all horizontally aligned pairs of upper face candidates at the distance *d(RER,LER)* with approximately equal number of edges. While searching for upper face candidates, the “primary” candidates were given the highest priority. Each candidate in the horizontal pair could have only one connection. In case of multiple connections, connections which had the longest length and the largest horizontal declination were eliminated until there was only one connection left.

2. Among the found upper face candidate pairs, there usually existed some noisy candidates, for example, elements of hair, closing, or decoration which needed to be eliminated. To do that, the distances between candidates and their relative locations in the face were verified.

    2.1. Although the algorithm was allowed to miss landmarks, however, at least one horizontal pair had to be found. If no pair was found, it was assumed that eye regions were merged together and edge x-projection was applied to the candidate with biggest number of edges located in the upper part of the image. After this step, edge map construction and edge orientation matching were applied to the received edge regions. After that the search started from the step 1.

    2.2. If one upper face candidate pair was found, it was labelled as eye region candidates. Using the eyebrow-eye distance, eyebrows above and eyes below the found pair location were searched for and if found any, the candidates were relabelled as eyes and eyebrows, respectively.

    2.3. If two upper face candidate pairs were found, they were verified by the eyebrow-eye distance and labelled as eyebrows and eyes, respectively.

    2.4. If more than two upper face candidate pairs were found, they were verified by the eyebrow-eye distance. If the distance between right or left candidates of the neighbouring pairs was less than the eyebrow-eye distance, the candidates were merged together. Otherwise, the candidate of the upper pair was eliminated. After there were one or two pairs left, the search started from the step 2.

    2.5. If the height of the upper face candidate was larger than a half of a dynamic parameter D_u that was calculated using a distance between mass centres of the upper face candidates, it was assumed that upper face candidate was merged with a hair. In this case, edge y-projection was applied in order to separate the candidate from the hair. After that the search started from the step 2.

3. At this step, the algorithm calculated a middle point between upper face landmark pair (eye or eye region pair) and built a vertical line of face symmetry called face axis.

4. The search for lower face landmarks was performed from top-to-bottom along the face axis. In order to verify spatial constraints between upper and lower face candidates and remove noisy candidates, the algorithm applied a dynamic parameter D_u. If the candidate mass centre was found down to the middle point between upper face landmark pair at a distance interval $[0.5D_u, 0.7D_u]$, it was marked as nose. If the candidate mass centre was found down to the midpoint at a distance interval $[1.2D_u, 1.7D_u]$, it was marked as mouth. This way, by utilizing geometrical constraints among the lower face candidates, the algorithm verified the upper face candidate locations.

    4.1. If only one lower face candidate was found, it was assumed that nose and mouth were combined together and edge y-projection was applied to separate these landmarks. After that the search started from the step 4.

Figure 4. Algorithm of structural correction of the located landmark candidates.

Fig. 2,*f* demonstrates the final result of the facial landmark localization. A localization result was defined as a rectangular box placed over the located region, not as a single point as typically has been the case. The location and size of the bounding box were calculated from the coordinates of the top, left, right, and bottom boundaries of the edge region. The mass centre of the located region indicated an estimate of the centre of the landmark.

IV. PERFORMANCE EVALUATION MEASURE

To our knowledge, the performance evaluation of different facial feature detectors proposed in the literature was given in terms of either visual inspection of the detection result or error measure calculated as a distance between manually annotated and automatically detected feature point locations. We do not consider visual inspection as an appropriate evaluation measure for any feature detector as it is a subjective decision and, therefore, can not give objective criteria of what to consider as a correct detection result. An error measure has been usually reported in terms of Euclidean pixel distance - the fewer pixels there were, the better the accuracy of the feature detector. In the work by Jesorsky et al. [20] a detection result was considered correct if the distance between manually annotated and automatically detected feature point location was less than 1/4 of the annotated intereye distance. This point measure is sufficient for all applications which can make use of a single pixel point result as an output of the feature detector. However, there is a number of applications which require a feature detector to find a region of facial landmark rather than to give a point solution. To our knowledge, there are no criteria for evaluation of the landmark detection result which is represented by a region in the image, not a single point.

In order to create a description of the landmark location in the image, we selected a set of characteristic points shown in Figure 5. Four points were selected to define locations of the eye, eye region, and mouth. These points defined the right, left, top, and bottom boundaries of these landmarks. Three points were used to describe locations of the eyebrow and nose. The eyebrow location was described by its top, bottom-left, and bottom-right points. Because it was unclear how to define the vertical dimensions of the lower nose, this region was defined by the centre point of the nose tip and locations of the nostrils. All the characteristic points were manually annotated in all the databases. Further, bounding boxes were built on the base of the selected characteristic points for each landmark in all the databases. The centre of the landmark was defined as the centre point of the bounding box [19].

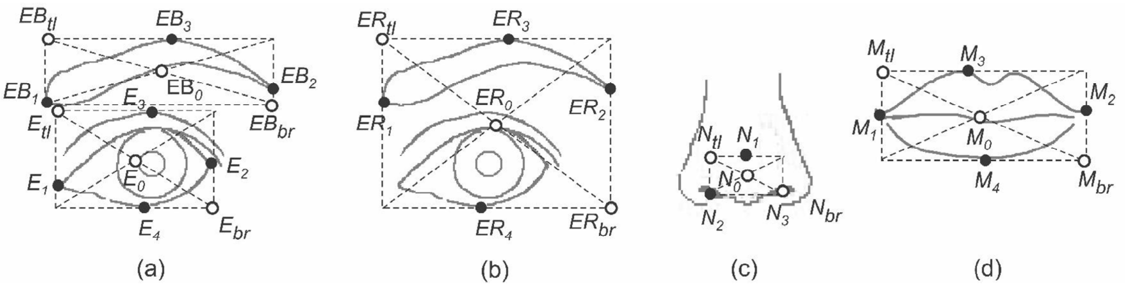


Figure 5. A set of characteristic points selected to define landmark location in the image: (a) eyebrow and eye; (b) eye region; (c) lower nose; (d) mouth. The bounding box that contains a landmark was defined by its top-left (*tl*) and bottom-right (*br*) bounding coordinates. The center point of the landmark was defined by the center point of the bounding box.

The centre point, top-left, and bottom-right coordinates of the bounding box were assumed to provide a good description of the landmark location in the image and be potentially useful for the purpose of the method performance evaluation. The pilot test was performed to verify this assumption. The characteristic points were manually annotated by 19 participants using a small dataset of images which well reflected the variations in facial expressions. The results of the pilot test supported our initial assumption; however, they also demonstrated that eyebrow characteristic points were difficult to define by humans and would not be as useful in the evaluation of the method performance as anticipated. In particular, some images contained subjects with very light eyebrows or eyebrows were covered by hair. For nearly all human subjects it was difficult to define boundary points for these landmarks. For some subjects, the difference in the annotation results was more than 15 pixels while the average length of the eyebrow was 38 pixels. As we aimed to compare the results of the automatic landmark localization to the human annotation results, the impact of poor definition of the eyebrow boundary points on the method performance evaluation was prevented by making the eyebrow a complementary landmark to locate. It means that the localization performance of the upper face landmarks depended only on the localization of the eyes.

The correctness of the localization result for nose was defined as a distance from manually annotated and automatically located centre of the nose tip. The correctness of the localization result for upper face landmarks and mouth was defined by a rectangular measure given in Equation 8:

$$\max(d(p_{tl},\overline{p}_{tl}),d(p_{br},\overline{p}_{br}))\le R\,,\ R = N\cdot StDev\,. \qquad (8)$$

where R is a performance evaluation measure; p_{tl}, p_{br}, $\overline{p}_{tl}$, and $\overline{p}_{br}$ define the centre point, top-left and bottom-right coordinates of the manually annotated and automatically located landmark, respectively; N is a number that sets a desirable accuracy of the localization result; $StDev = 2$ pixels is a standard deviation of the manual annotation averaged over all the characteristic points in the pilot test. If $\overline{p}_{tl}$ and $\overline{p}_{br}$ were found inside the manually annotated landmark position, $\overline{p}_{tl}$ and $\overline{p}_{br}$ should be located in the top-left and bottom-right quadrants of the bounding box which includes the annotated landmark.

The rate of the landmark localization was defined as a ratio between a total number of correctly located landmarks and a total number of images used in testing (as there was one face per image). Eye region localization was counted correct in both cases – if bounding box included both eye and eyebrow, or if eye and eyebrow were located separately. If eyebrow was located as a separate region, it was obligatory that a corresponding eye was also found. Localization result was considered as a misclassification if landmark was located correctly but erroneously classified as another landmark. Localization result was classified as wrong if bounding box covered several neighbouring facial landmarks, excluding the case of eyes and eyebrows located as one region. Localization result was counted as a false localization if bounding box included a non-landmark, for example, wrinkles in the face, ears, clothing, hair, and eyebrow located without a corresponding eye.

V. RESULTS

Figure 6 demonstrates average orientation portraits of facial landmarks for all the databases. The results confirmed that individual orientation portraits of facial landmarks generally followed the rules predefined by the orientation model. On the contrary, the orientation portraits of noisy regions usually had an arbitrary distribution of local oriented edges and were discarded by the orientation model. Figures 7 and 8 demonstrate the final results of the landmark localization on the Cohn-Kanade and JAFFE databases, respectively. A bounding box that was built on the basis of the top-left and bottom-right coordinates of the edge region defined the location of the landmark in the image. The mass centre of the edge region indicated an estimate of the centre of the landmark. As figures show, the shape and size of the landmarks varied significantly with changes in facial expressions. Accordingly, the size of the bounding box was dynamic and varied in compliance with a size of the located edge region. This allowed, for example, to locate open and tight mouth as illustrated in Figures 8b and g.

Figure 9 demonstrates the results of the method performance evaluated by the conventional point measure calculated as a distance between manually annotated landmark centre and mass centre of the

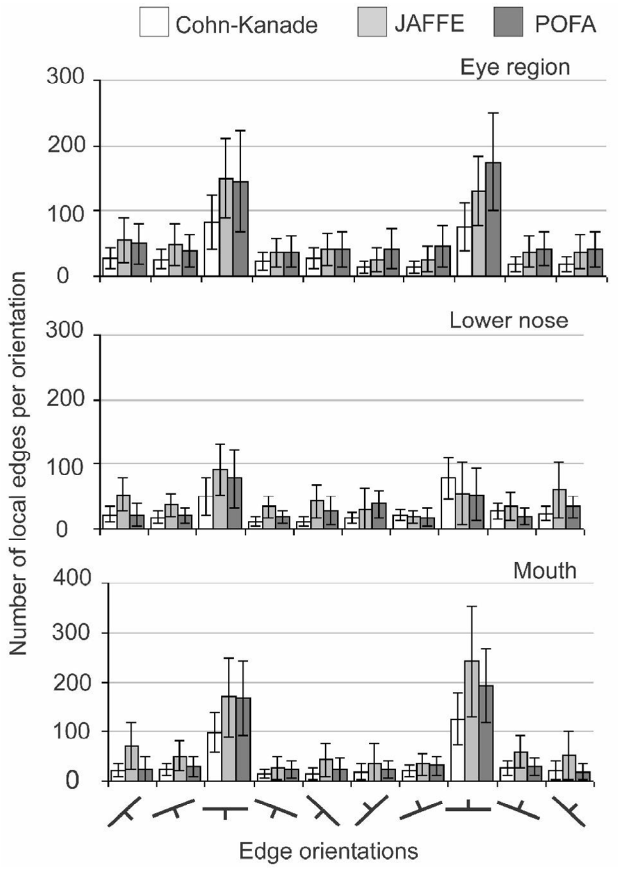
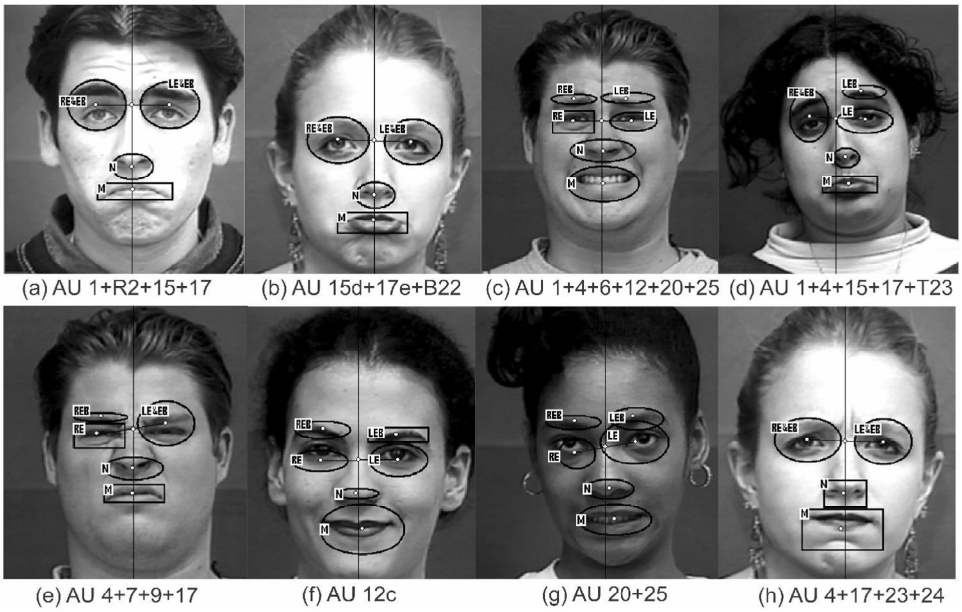


Figure 6. Examples of landmark orientation portraits averaged over all datasets. The error bars show plus/minus one standard deviation from the mean values.

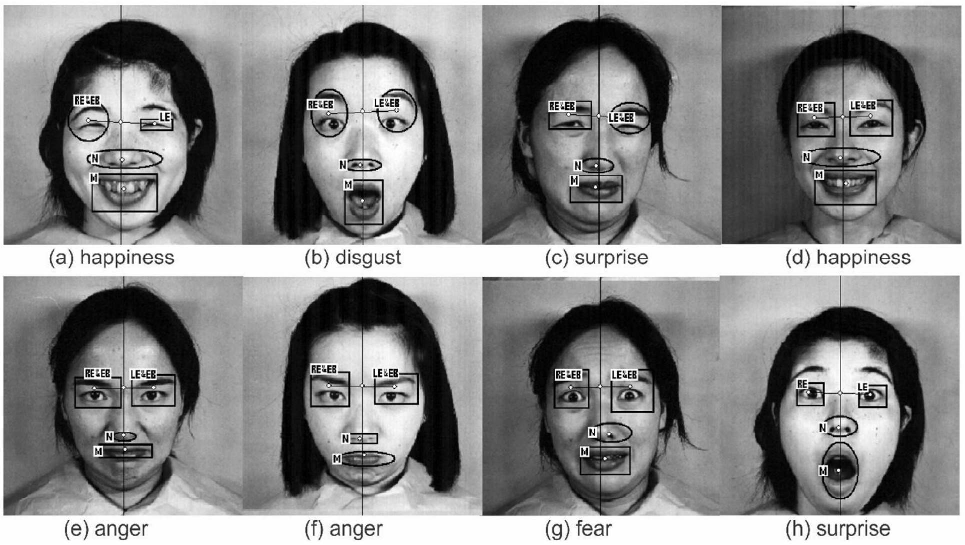
automatically located edge region. The results of the method performance evaluated by the proposed rectangular measure are shown in Figure 10. The performance of the method in this case was evaluated by the rectangular measure that was calculated as a maximum distance between manually annotated and automatically located top-left and bottom-right corners of the rectangular box which contained a landmark (Equation 8). In both cases, as a measure of distance between annotated and automatically found landmark location we used a standard deviation of the manual annotation of the characteristic points averaged over all subjects in the pilot test.

Figures 9 and 10 show that for Cohn-Kanade datasets the performance of the method was higher for distance measure R_1 as compared to R_2 which equalled to 1/5 and 1/4 of the average intereye distance for a given dataset, respectively. For POFA and JAFFE neutral and expressive datasets, the landmark localization rates increased rapidly with the increase in the distance measures. In this case, the method located landmarks with relatively high rates for both distance measures.


(a) AU 1+R2+15+17 (b) AU 15d+17e+B22 (c) AU 1+4+6+12+20+25 (d) AU 1+4+15+17+T23

(e) AU 4+7+9+17 (f) AU 12c (g) AU 20+25 (h) AU 4+17+23+24

Figure 7. Examples of the final localization results with “primary” landmarks (rectangles) and “secondary” landmarks (ovals) of the landmark localization in the Cohn-Kanade images. Reprinted with permission.


(a) happiness (b) disgust (c) surprise (d) happiness

(e) anger (f) anger (g) fear (h) surprise

Figure 8. Examples of the final localization results with “primary” landmarks (rectangles) and “secondary” landmarks (ovals) of the landmark localization in the JAFFE images. Reprinted with permission.

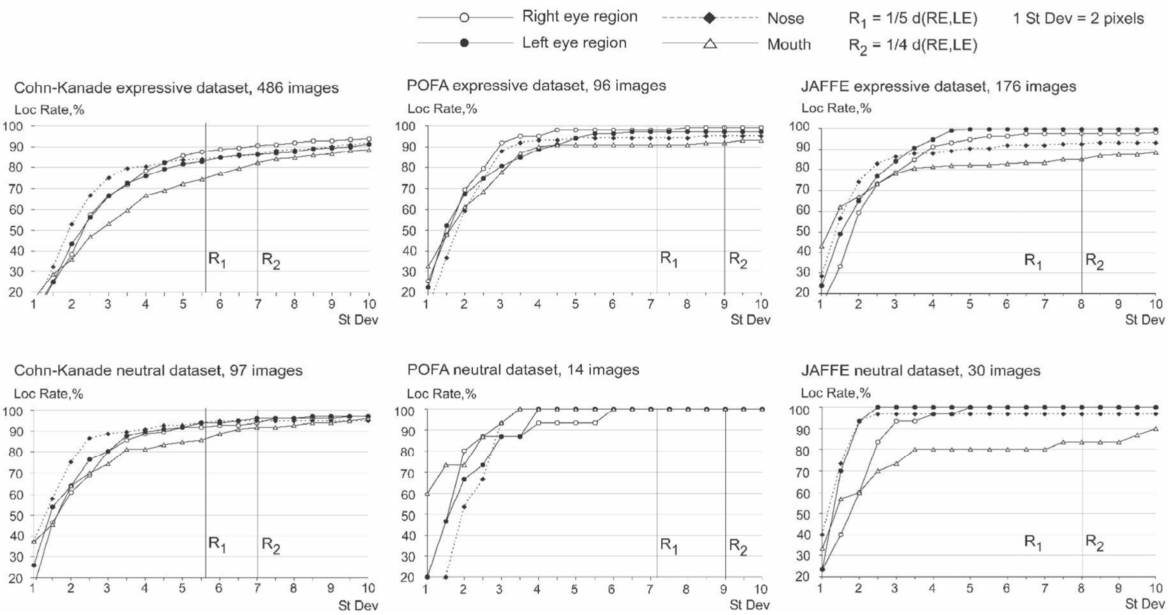


Figure 9. The performance of the method on Cohn-Kanade (first column), POFA (middle column), and JAFFE (right column) neutral and expressive datasets evaluated by the conventional point measure. The vertical lines indicate a distance measures which equaled to 1/4 and 1/5 of the average intereye distance for a given dataset.

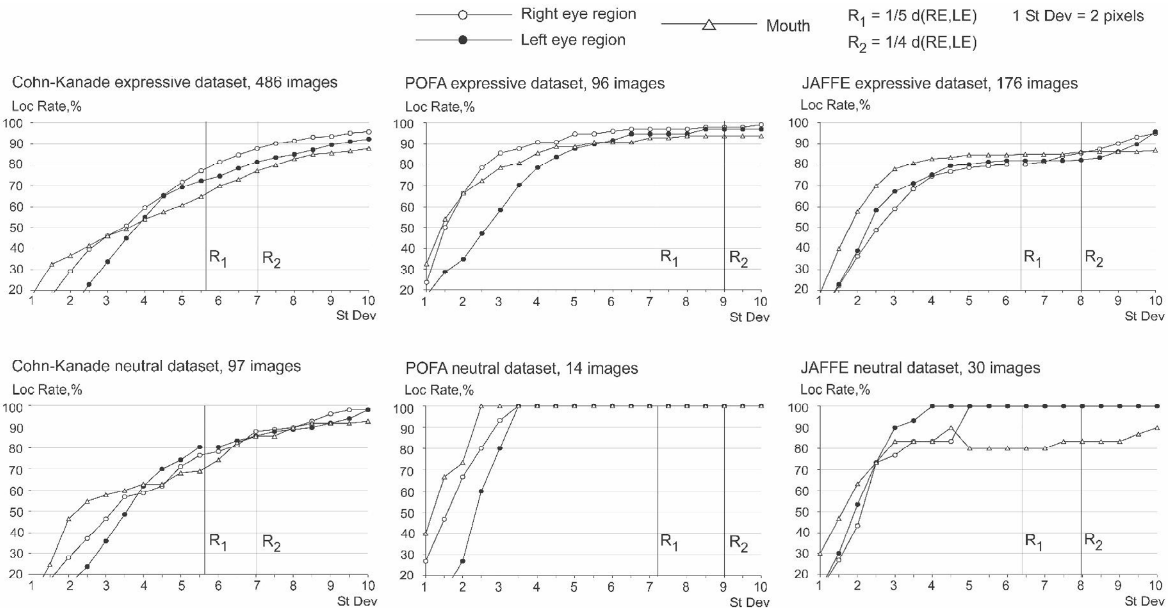


Figure 10. The performance of the method on Cohn-Kanade (first column), POFA (middle column), and JAFFE (right column) neutral and expressive datasets evaluated by the proposed rectangular measure. The vertical lines indicate a distance measures which equaled to 1/4 and 1/5 of the average intereye distance for a given dataset.

The performance of the method evaluated by the conventional and rectangular evaluation measures is summarized in Table II. For both types of the evaluation criteria the landmark localization was robust with respect to facial expressions as landmarks were located with nearly equal rates from neutral and expressive datasets. A decrease in the localization rates was observed for Cohn-Kanade and JAFFE expressive datasets; on the whole, however, the overall performance of the method as evaluated by the point and rectangular measures was high. A good method performance as evaluated by the rectangular measure reflected the fact that in most cases the method located landmark positions precisely meaning that inside the bounding box area surrounding landmark was less than the actual size of the landmark (for an opposite example refer to Figure 7h).

TABLE II: METHOD PERFORMANCE ON NEUTRAL AND EXPRESSIVE DATASETS, $R = \frac{1}{4} \mathrm{d}(\mathrm{RE}, \mathrm{LE})$

Datasets	Method performance evaluated by point measure				Method performance evaluated by rectangular measure		
	L eye r	R eye r	Nose	Mouth	L eye r	R eye r	Mouth
Cohn-Kanade, expressive	90%	86%	86%	82%	88%	81%	77%
Cohn-Kanade, neutral	94%	96%	95%	92%	88%	86%	86%
POFA, expressive	99%	97%	95%	92%	98%	97%	94%
POFA, neutral	100%	100%	100%	100%	100%	100%	100%
JAFFE, expressive	97%	99%	92%	85%	86%	82%	86%
JAFFE, neutral	100%	100%	97%	83%	100%	100%	83%
Average	94%				91%		

The errors in the landmark localization were tracked manually and classified into misclassification, false localization, and wrong localization groups (Table III). Misclassification errors appeared on the last stage of the method due to errors in the algorithm of structural correction. Misclassification of the upper face landmarks was mainly due to the classification of eyebrows and eye regions as eyes. Lower face misclassification was due to the classification of nose as mouth and vice versa. Wrong localizations were mainly observed in the localization of lower face landmarks. As it was expected, it occurred mainly due to the effect of lower face AUs 9, 10, 11, and 12 activated alone or in combinations with other AUs, for example, in expressions of anger, disgust, and happiness. AUs 4, 6, and 7 sometimes caused the merging of the upper face landmarks into one region. Figure 11 illustrates some examples of the localization errors.

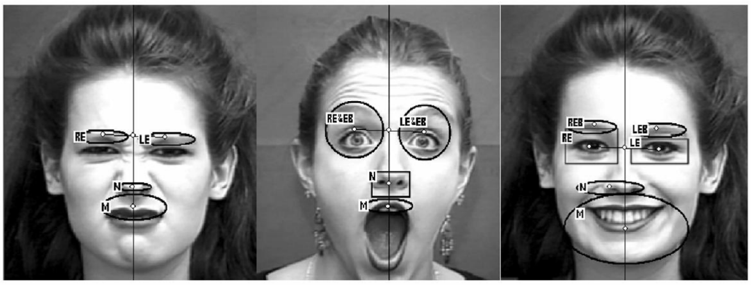
TABLE III: SUMMARY OF ERRORS OF THE METHOD PERFORMANCE ON NEUTRAL AND EXPRESSIVE DATASETS

Datasets	Misclassifications		Wrong localizations		False localizations	
	Upper face landmarks	Lower face landmarks	Upper face landmarks	Lower face landmarks	Upper face landmarks	Lower face landmarks
Cohn-Kanade, expressive	34	44 (7)	7	15	50	28
Cohn-Kanade, neutral	3	2 (1)	2	-	7	3
POFA, expressive	5	2	-	4	-	3
POFA, neutral	-	-	-	-	-	-
JAFFE, expressive	3	1 (3)	1	5	-	4
JAFFE, neutral	-	-	-	-	-	-

Note that numbers in brackets define wrong landmark localizations which resulted into the misclassification error.

It has been demonstrated that the method produced the lowest localization rates and biggest number of localization errors for Cohn-Kanade expressive dataset. This fact suggested a further analysis of the method performance on this database to reveal the influence of particular facial behaviours on the localization results as evaluated by the rectangular measure. Additionally, we were interested what kind of improvement in the method performance was achieved as compared to the earlier version of the method [14]. In that study we discovered that the method performance was especially deteriorated by the lower face AUs 9, 10, 11, and 12 and some combinations of these AUs with others. A comparison of the landmark localization rates of the present and earlier versions of the method was made for images with upper face AUs and AU combinations (Table IV), lower face AUs (Table V), and lower face AU combinations (Table VI).

As it is seen from the tables, localization rates were increased over 20% for 25 out of 47 different facial deformations. A significant improvement in the landmark localization was achieved for images with



(a) AU 4+6+7+9d+17d+25 (b) AU 1+2+5d+25+27 (c) AU 6+12+25

Figure 11. Examples of errors of the landmark localization: (a) eyebrows were misclassified as eyes; (b) upper lip was classified as mouth; and (c) mouth was merged with chin. Images are courtesy of the Cohn-Kanade and JAFFE databases. Reprinted with permission.

lower face AUs and AU combinations. For example, the improvement in the localization rates for lower face AUs 9, 9+17, 12+16, 16+20 was over 30%. The same improvement was achieved for upper face AU combinations 1+6, 4+6, 6+7.

Due to the fact that some AUs were not presented in the dataset or the number of images was too few (less than 6), only a limited number of AUs and AU combinations was used. This type of the result classification allowed the results to belong to more than one group. Moreover, AUs from the tables

TABLE IV: IMPROVEMENT OVER UPPER FACE AUS AND AU COMBINATIONS

Study / *AUs*	1	2	4	5	6	7	43/45	1+2	1+4	1+5	1+6	1+7	2+4	2+5	4+5	4+6	4+7	4+43/45	6+7
Ave Loc Rate, % Present study	84	83	85	86	77	81	91	83	82	84	92	80	80	84	85	82	82	84	76
Ave Loc Rate, % Study [14]	81	86	64	84	52	61	63	86	73	86	62	58	82	86	72	40	62	56	46
Improvement, %	2	-3	21	2	25	20	28	-3	9	-2	30	21	-2	-2	13	42	20	28	30

Note that AU43 (eye closure) and AU45 (blink) were combined together because they both have the same visual effect on the facial appearance and different durations of these AUs can not be measured from the static images.

TABLE V: IMPROVEMENT OVER LOWER FACE AUS

Study / *AUs*	9	10	11	12	14	15	16	17	20	23	24	25	26	27
Ave Loc Rate, % Present study	84	83	85	77	68	94	88	89	83	88	86	84	92	83
Ave Loc Rate, % Study [14]	30	58	66	60	77	70	67	64	68	78	79	68	71	90
Improvement, %	53	25	20	17	-9	24	21	25	16	9	7	16	21	-6

TABLE VI: IMPROVEMENT OVER LOWER FACE AU COMBINATIONS

Study / *AUs*	9+17	11+20	11+25	12+16	12+20	12+25	16+20	16+25	17+23	17+24	20+25	23+24	25+26	25+27
Ave Loc Rate, % Present study	83	88	88	86	63	74	94	86	87	86	83	86	91	84
Ave Loc Rate, % Study [14]	41	67	67	50	42	54	54	61	72	76	67	77	70	90
Improvement, %	42	21	21	36	21	19	40	25	14	9	16	10	21	-6

usually occurred singly or in conjunction with other AUs which are not represented in the tables. That is why the present results revealed an indirect effect of different AU and AU combinations on the landmark localization.

VI. DISCUSSION

A fully automatic method was designed for feature-based localization of facial landmarks in static grey-scale images. The local oriented edges served as basic image features for expression-invariant representation of facial landmarks. The results confirmed that in the majority of the images, the orientation portraits of facial landmarks had the same structure as predefined by the orientation model. This allowed to discard non-landmark regions like, for example, wrinkles in the face, ears, earrings, and elements of hair and clothing. Structural correction of the located candidates based on face geometry model further improved the overall performance of the method.

The performance of the method was examined on three databases of facial images showing complex facial expressions. The complexity of the expressions was presented by a variability of the deformations in soft tissues (wrinkles and protrusions), variety of mouth appearances including open and tight mouth, visible teeth and tongue, self occlusions (semi- and closed eyes and bitted lips). The results of the landmark localization were evaluated by the conventional point measure and the proposed rectangular measure that represented a localization result as a rectangular box, not a single point. Both types of evaluations demonstrated a sufficiently high overall performance of the method (94% as evaluated by the point measure and 91% as evaluated by the rectangular measure) that is comparable to some extent with performance of other facial feature detectors reported in the literature [9,10,21,22,23].

The comparison of the present results with the results of the method testing on the Cohn-Kanade dataset obtained in the earlier study [14] demonstrated that the modifications made to the method significantly improved the overall performance of the method. One has to be cautious when comparing the localization rates from Tables IV-VI because the implementation of the method, test setup, and evaluation criteria were different in these two studies. However, the general trend in the method improvement can be brought into the light.

As one significant problem in the earlier method was the deteriorating effect of expressions of happiness (when AUs 6 and 12 are usually activated), anger (when AUs 4, 6, and 7 are usually activated), and disgust (when AUs 9, 10, and 11 are usually activated). Occurring along or in conjunction, the listed

AUs caused a wrong localization error in the landmark localization. In the current version of the method, the applied procedure of x/y-edge projection in many cases allowed the merged landmarks to be separated. In particular, the merged nose-and-mouth landmark was successfully separated, especially in the images with AUs 10, 11, 12, 15, 16, 17, 21, 11+20, 11+25, 12+20, 16+25, and 25+26. Further, an essential improvement in locating landmarks was achieved for images with AUs 4, 6, 7, 43/45, 1+6, 1+7, 4+6, 4+7, 4+43/45, 6+7, and 9. These facial behaviours typically cause the whites of the eyes to become less or not at all visible in the face. These facial behaviours cause serious errors in the performance of eye detectors which rely particularly on the searching for pupils and whites of the eyes in the image. The present method can deal with self-occlusions in the eye regions by analyzing a local edge description of the whole eye region.

Generally, in all the databases the mouth region demonstrated a greater variability in its appearance than the regions of eyes or nose. For example, in the images of surprise and happiness the mouth appearance usually was represented by open mouth (when AU combination 25+26 is activated) sometimes with visible teeth and tongue. In the images of anger the mouth could be opened (AU 22+25+26) or closed with tightened lips (AU 23), pressed lips (AU 24), and even lips sucked inside the mouth (AU 28) so that the red part of the mouth became not visible in the face. These facial behaviours restrict the applicability of the colour-based methods of mouth detection in which colour is the key feature in the mouth representation. On the contrary, the proposed method was able to find the mouth regardless of whether the mouth was open or closed and whether the lips, teeth or tongue were visible or not (Figures 7-8).

Emphasizing the simplicity of the proposed method, we conclude that it can be used in the preliminary localization of regions of facial landmarks for their subsequent processing where coarse landmark localization is followed by fine feature detection. For example, eye and mouth corners can be searched for in the located regions. The method can be applied directly to the image without any image alignment given that a face takes the biggest part of the image. The method does not require intensive training either. These qualities gives the method an advantage over PCA, AdaBoost, or neural network based methods which require either facial alignment or intensive training made preliminary to the facial landmark localization.

Currently, the method was applied to static images where no temporal information was available, only structural information was used. The processing time of the method was slowed down by the stage of edge detection where each point of the image was checked for a local oriented edge. To increase the speed of the method for practical applications, face region estimation can be utilized as a preliminary step to edge extraction in order to discard those parts of the image which a priori can not contain facial landmarks. Our future plan is to utilize the developed method in real-time system of facial landmark detection and tracking for the purpose of facial expression analysis in human-computer interaction.

ACKNOWLEDGMENTS

This work was supported by the Finnish Academy (project number 177857 and 1115997), the University of Tampere, and the Tampere Graduate School in Information Science and Engineering. The authors thank the creators of Cohn-Kanade and JAFFE databases for the permission to reprint examples of the expressive images.

REFERENCES

[1] P. Ekman, The argument and evidence about universals in facial expressions of emotion, in: H. Wagner, A. Manstead (eds.), Handbook of Social Psychophysiology, Lawrence Associates Press, London, 1989, pp. 143-164.

[2] P. Ekman, W. Friesen, Facial action coding system (FACS): A technique for the measurement of facial action, Consulting Psychologists Press, Palo Alto, California, 1978.

[3] P. Ekman, W. Friesen, J. Hager, Facial action coding system (FACS), A Human Face, Salt Lake City, UTAH, 2002.

[4] E. Hjelmas, B. Low, Face detection: A survey, Computer Vision and Image Understanding 83 (2001) 235–274.

[5] M. Yang, D. Kriegman, N. Ahuaja, Detecting face in images: A survey, IEEE Trans. Pattern Analysis and Machine Intelligence 24 (2002) 34-58.

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

[7] M. Burl, T. Leung, P. Perona, Face localization via shape statistics, in Proc. 1st Int. Workshop Automatic Face and Gesture Recognition, Zurich, Switzerland, June 1995, pp. 154-159.

[8] L. Wiskott, J.-M. Fellous, N. Kruger, C. von der Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Analysis and Machine Intelligence 19, 7 (1997) 775–779.

[9] R. Feris, J. Gemmell, K. Toyama, V. Krüger, Hierarchical wavelet networks for facial feature localization, in Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition, Santa Barbara, CA, May 2002, pp. 118-123.

[10] D. Cristinacce, T. Cootes, Facial feature detection using AdaBoost with shape constraints, in Proc. 14th British Machine Vision Conf., Norwich, England, September 2003, pp. 231-240.

[11] P. Viola, M. Jones, Robust real-time face detection. Int. J. Computer Vision 57, 2 (2004) 137-154.

[12] Y. Gizatdinova, V. Surakka, Feature-based detection of facial landmarks from neutral and expressive facial images, IEEE Trans. Pattern Analysis and Machine Intelligence 28, 1 (2006) 135-139.

[13] I. Guizatdinova, V. Surakka, Detection of facial landmarks from neutral, happy, and disgust facial images, in Proc. 13th Int. Conf. Central Europe on Computer Graphics, Visualization, and Computer Vision, Plzen, Czech Republic, January 2005, pp. 55-62.

[14] Y. Gizatdinova, V. Surakka, Effect of facial expressions on feature-based landmark localization in static grey scale images, Int. Conf. Computer Vision Theory and Applications, Madeira, Portugal, January 2008, pp. 259-266.

[15] Y. Gizatdinova, V. Surakka, Automatic detection of facial landmarks from AU-coded expressive facial images, Int. Conf. Image Analysis and Processing, Modena, Italy, September 2007, pp. 419-424.

[16] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition, Grenoble, France, March 2000, pp. 46-53.

[17] P. Ekman, W. Friesen, Pictures of facial affect, Consulting Psychologists Press, Palo Alto, California, 1976.

[18] M.J. Lyons, Sh. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with Gabor wavelets, in Proc 3d IEEE Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, April 1998, pp. 200-205.

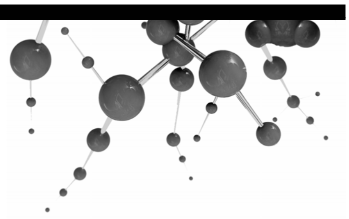
[19] L. Farkas, Anthropometry of the head and face, (2 ed), Raven, New York, 1994.

[20] O. Jesorsky, K.J. Kirchberg, R.W. Frischholz, Robust face detection using the hausdorff distance, Lecture Notes in Computer Science, Proc. of the 3d Int. Conf. Audio- and Video-Based Person Authentication, Halmstad, Sweden, June 2001, pp. 90–95.

[21] P. Campadelli, R. Lanzarotti, G. Lipori, E. Salvi, Face and facial feature localization, in Proc. 13th Int. Conf. Image Analysis and Processing, Gagliari, Italy, September 2005, pp. 1002-1009.

[22] B. Fröba, C. Küblbeck, Orientation template matching for face localization in complex visual scenes, in Proc. Int. Conf. Image Processing, Vancouver, September 2000, pp. 251-254.

[23] D. Shaposhnikov, A. Golovan, L. Podladchikova, N. Shevtsova, X. Gao, V. Gusakova, Y. Gizatdinova, Application of the behavioural model of vision for invariant recognition of facial and traffic sign images, J. Neurocomputers: Design and Application 7-8 (2002) 21-33, (in Russian).



Publication VI

Erola J., Gizatdinova Y., and Surakka V. (2008). Automatic Real-Time Localization of Frowning and Smiling Faces under Head Pose Variations. *Proceedings of International Conference on Signal Processing, Computational Geometry and Artificial Vision (ISCGAV'08)*, WSEAS Press, 22-27. (The extended version of this paper was published in *WSEAS Transactions on Signal Processing*, 1 (4), 463-473, 2008.)

Available online at:
http://www.wseas.org (requires subscription)

Automatic Real-Time Localization of Frowning and Smiling Faces under Head Rotation Variations

JOUNI EROLA, YULIA GIZATDINOVA, AND VEIKKO SURAKKA
Department of Computer Sciences
University of Tampere
Kanslerinrinne 1, 33014
FINLAND
jouni@tuu.fi, {yulia.gizatdinova, veikko.surakka}@cs.uta.fi

Abstract: - A new method for real-time face localization from a streaming color video was developed. The method consisted of three stages. First, the face-like skin-colored image region was segmented from the background and transformed into the grey scale representation. Second, the cropped image was convolved with Sobel operator in order to extract local oriented edges at 16 orientations. The extracted local oriented edges were grouped together to form regions of interest which represent landmark candidates. Further, the candidates were matched against edge orientation model to verify the existence of the landmark in the image. Finally, the located landmarks were next spatially arranged into the face–like constellations. The best face-like constellation of the landmark candidates was defined by a new scoring function. The test results demonstrated that the proposed method located expressive faces with high rates in real time from facial images under controlled head rotation variations.

Key-Words: - Face localization, Facial landmarks, Sobel edge detection, Frontal-view geometrical face model, Facial expression, Head rotation.

1 Introduction

In automatic face detection, different feature detectors are applied in order to find a face-like region in static images or video frames. If it is known in advance that face is shown in the image, the task comes to find a true face location. The found face location is further delivered as an input to various systems of automatic face analysis such as face and facial expression recognition and perceptual vision-based user interfaces. To ensure that these systems work in real time efficiently and robustly, face detection is aimed to provide high speed and accuracy of the detection process.

The main challenge in face detection is to find a face representation that remains robust with respect to various changes in facial appearance since face varies noticeably with changes in environmental conditions (e.g. illumination, out- and in-plane head rotations, scene complexity, resolution, occlusions, etc.), ethnicity (i.e. skin color), gender, and facial expressions (i.e. emotional and social signals in the face). Many attempts have been undertaken to automatically detect faces from static images and video [10,22]. We can roughly classify the existing techniques of face detection as belonging to appearance- or feature-based approaches. The appearance-based approach uses holistic features of the image and considers a face as a whole. Methods of Principle Component Analysis (PCA) have been widely adopted for the purpose of face detection [9,21]. Generally, PCA-based methods can handle faces with nearly the same pose, constant illumination, and moderate facial expressions. Apart from PCA-based methods, there are also learning-based methods like boosted classifiers [19], support vector machines [12], and neural networks [14] which have to be trained on the representative sets of face and non-face images.

The feature-based approach to face detection utilizes local features of the image. This approach can overcome the constraints placed by illumination change, head rotations, and facial expressions. This is due to the fact that it is based on modeling local texture information around individual facial landmarks and modeling global shape information on spatial arrangement of the located landmark candidates. Facial landmarks are typically those which are the most distinctive for humans - eyebrows, eyes, nose, and mouth. These landmarks encode critical information on facial expressions and head movements that is used in automatic face analysis. In practice, feature-based face detection includes a selection of feature representation and a design of feature detector. Different features can be detected from the image or video frame, for example, edges, colors, points, lines, contours, etc.

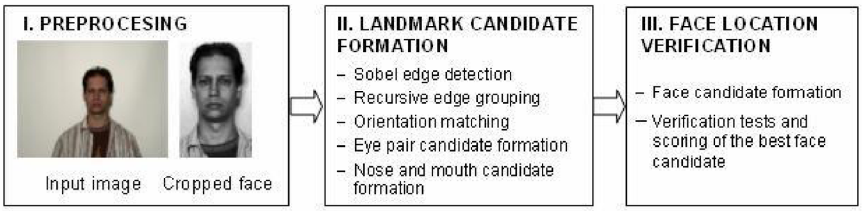

I. PREPROCESING
Input image
Cropped face
II. LANDMARK CANDIDATE FORMATION
– Sobel edge detection
– Recursive edge grouping
– Orientation matching
– Eye pair candidate formation
– Nose and mouth candidate formation
III. FACE LOCATION VERIFICATION
– Face candidate formation
– Verification tests and scoring of the best face candidate

Figure 1. Data flow in our system of real-time face localization.

These features provide a meaningful and measurable description of the face as they represent specific visual patterns used to identify corresponding structures between images. Extensive work has been focused on shape representation of facial landmarks [2,3,4,15]. Many proposed face detectors utilize edges [6], grey-scale values [17], and their combinations [16]. A multi-resolution and multi-orientation representation of the image has been widely adopted for the purpose of landmark detection and demonstrated to be effective in face detection under expression and small pose variations [7,20,23]. However, these methods are generally computationally expensive and, therefore, are not applicable for the task of real-time face and facial feature detection.

In this paper, we introduce a feature-based method of face localization from streaming color video. It is based on the multi-orientation image representation that helps in composing facial landmarks. This is followed by spatial arrangement of the located landmark candidates. We present a new scoring function designed to define the best face-like constellation of the candidates. The landmarks to be located in the face are centers of the eyes, nose tip, and center of the mouth. In the subsequent sections we demonstrate how our proposed method successfully overcomes speed limitations of the feature-based face detection methods.

2 Face Localization

The method for real-time face localization consists of three stages depicted in Fig. 1. Given a facial video, the first stage finds a rough approximation of the face location in each video frame. The second stage forms and extracts all possible landmark candidates from the cropped face region. The last stage decides whether constellations formed from the located landmark candidates meet requirements placed by the geometrical frontal-view facial model, and if so, defines the location of the best-scored face candidate. Below we explain each stage of the method in more detail.

2.1 Preprocessing

On this stage, a face-like image region is segmented from the background. To do that we first apply the procedure of histogram equalization to each video frame for all three RGB color channels. This calculation is not computationally expensive and is widely used to allow areas of low local contrast to gain a higher contrast without affecting the global contrast of the whole image [8]. After this, the original RGB image is transferred into YCbCr chromatic color space as proposed in [11].

Next, we segment the skin-colored regions of the image similarly to the method proposed in [1]. This procedure allows for noisy regions of the image to be discarded at the early stage of the processing and, therefore, focuses the following stages of the method on those parts of the image in which the face is more likely to be located. The Gaussian-fitted skin color model is used for this purpose. The idea that lies behind is that a distribution of skin color for different people is clustered in the chromatic color space and can be represented by a Gaussian model. The likelihood P of a skin color for any pixel *(x,y)* of the image thus can be obtained with a Gaussian-fitted skin color model:

$$P = e^{\left(-\frac{1}{2\cdot Cv}\cdot\left((Cb-Cb_{mean})^2+(Cr-Cr_{mean})^2\right)\right)}, \quad (1)$$

where Cv is a covariance matrix; Cb is a blue chromatic value; Cr is a red chromatic value; Cb_{mean} is an average blue chromatic value; and Cr_{mean} is an average red chromatic value.

The skin-colored regions are next segmented from the rest of the image through a thresholding process. The parameters for a thresholding are selected experimentally using a small image set from the database. Finally, the received regions are cropped from the background and the procedure of histogram equalization is applied for them. This stage outputs 8-bit grey scale image of the face. If a face is not extracted, the following steps of the method are applied to the whole image converted into 8-bit grey scale representation.

3.2 Landmark Candidate Formation

On the next stage, the cropped face image is convolved with Sobel operator [8] in order to extract local oriented edges at 16 orientations. The edge points are further recursively grouped together to form regions of interest which represent candidates for facial landmarks. The shapes of the bounding boxes which are placed over the located regions of interest are analyzed next. If a candidate is bounded by a box which has height much bigger than its width, the candidate is eliminated. After this, among the located candidates there still exist many noisy regions which have to be eliminated. In order to define regions which represent landmark candidates we analyze local properties of the located regions of interest. As it has been demonstrated earlier [7], regions of facial landmarks have a characteristic distribution of local oriented edges with two horizontal dominants (Fig. 2*a* and 2*b*). On the other hand, non-landmark regions which are, for example, elements of face, hair, clothing, and decoration typically do not have a characteristic structure of the oriented edges. These regions demonstrate a random distribution of the oriented edges (Fig. 2*c* and 2*d*). This local property of the located regions of interest allowed us to discard noisy regions while preserving regions which contain facial landmarks.

In order to classify the located candidates and find their proper spatial arrangement, the proposed method applies a set of verification rules which are based on face geometry. The knowledge on face geometry is taken from the anthropological study by Farkas [5]. This thorough study examined thousands of Caucasians, Chinese, and African-American subjects in order to determine characteristic measures and proportion indexes of the human face. We performed several tests to verify the existing facial measures, calculate new measures, and built a frontal-view geometrical face model depicted in Fig. 3. The anthropometric facial features and measures of the model are described in Table 1. Center points of facial landmarks are calculated as mass centers of the located edge regions. It has been demonstrated that facial measures from the table can slightly vary between subjects of different gender, age, and race [5]. Therefore, we define constrains for facial measures as intervals between minimum and maximum values for a given measure. All distance constrains from the table are defined as percentages of the interocular distance $d(E_R, E_L)$. As our preliminary tests demonstrated, this set of facial measures and their corresponding constrains achieved good results in composing face-like constellations from the located landmark candidates.

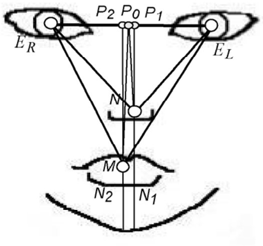
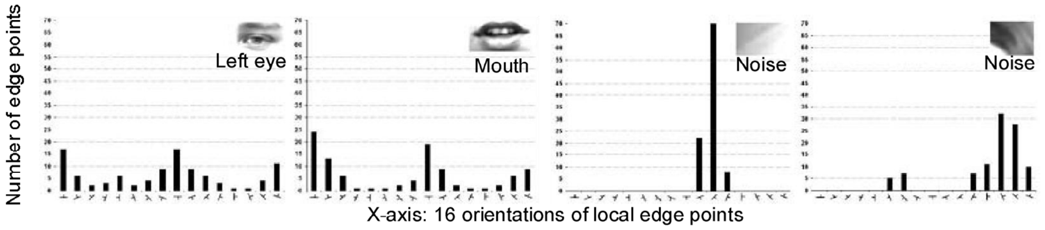


Figure 3. Frontal-view geometrical face model.

The classification of the landmark candidates proceeds as follows. The eye pair candidates are found first as any two candidates aligned nearly horizontally. After this step, all possible nose and mouth candidates for each found eye pair candidate are independently searched for using constrains of the frontal-view geometrical face model.

3.3 Face Location Verification

On the last stage, the defined eye-nose and eye-mouth candidates are combined together into a complete face candidate so, that the eye pair is the same for both eye-nose and eye-mouth candidates. Each found face candidate consisting of four facial landmarks is given a score. A score is calculated as a sum of intermediate scores which show how well a face candidate performs verification tests. The verification tests are fuzzy rules defined as follows: *Test 1* checks the horizontality of the eye pair candidate. The eye pair candidate that has the most horizontal position in the image as compared to



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

Figure 2. Facial landmarks with a characteristic distribution of the oriented edges (*a* and *b*) and noisy regions with a random distribution of the oriented edges (*c* and *d*).

Table 1: Features and measures of the frontal-view geometrical face model (and their constrains).

Feature/measure	Feature description
E_R , E_L , N , M	Centers of the landmarks
$d(E_R,E_L)$	Interocular distance
N_1, N_2	Perpendiculars to $d(E_R,E_L)$
P_1 , P_2	Cross points of $d(E_R,E_L)$ and N_1 and N_2, correspondently
P_0	Middle point between P_1 and P_2
$d(N,M)$	Nose-mouth distance (30-110% of $d(E_R,E_L)$)
$\angle NP_0M$	Nose-mouth angle (0-16°)
$\angle E_RNP_1$, $\angle E_LNP_1$ $\angle E_RMP_2$, $\angle E_LMP_2$	Eye-nose and mouth-eyes angles (0-13°)
$d(E_R,N)$, $d(E_L,N)$	Eye-nose distance (25-120% of $d(E_R,E_L)$)
$d(E_R,M)$, $d(E_L,M)$	Eye-mouth distance (60-160% of $d(E_R,E_L)$)

others gives the lowest score for a given face candidate. *Test 2* checks angles $\angle E_RNP_1$, $\angle E_LNP_1$, $\angle E_RMP_2$, and $\angle E_LMP_2$ - face candidate with small angles gets low scores, and vice versa. *Test 3* considers the result of the previous face localization. If the landmark center points in the previous frame are nearly the same as compared to those in the current frame (face is not moving), a face candidate gets a score which is lower than in the opposite case. *Test 4* checks sizes of the located landmark candidates – the biggest values give the lowest score for a face candidate. *Test 5* checks widths of the landmark candidates in the facial configuration. It has been validated that mouth is usually wider than eyes and nose has nearly the same width as eyes have [5]. The closer face candidate satisfies to this criterion, the lower score it gets from the test. *Test 6* utilizes the property of face symmetry and checks sizes of both eyes. If eyes have the same size, a face candidate gets the lowest score from this test. Each verification test is also given a weight. We performed a number of tests to define optimal weights for each test. This way, each test gives as its output a relative score for a given face candidate. In order to select the best-scored face-like constellation of the located landmarks, a new scoring function is introduced:

$$P_r = MAX - \frac{MAX}{P_{cand}/P_{\min}} \tag{2}$$

where *MAX* is a maximum score for a given face candidate (we used 100); P_{cand} is a current score for a given face candidate; $P_{\min}$ is the lowest score achieved among all face candidates. This way, if we have several face-like constellations of the landmarks, we select the one that gives the highest score P_r.

A localization result is considered correct if a distance between manually annotated and automatically located landmark location met the requirement placed by a performance evaluation measure elaborated in [13]:

$$d_{eye} = \frac{\max\left(d(E_R,E'_R),d(E_L,E'_L)\right)}{d(E_R,E_L)} \tag{3}$$

where $d(a,b)$ is Euclidean distance between point locations *a* and *b*; E_R , E_L are manually annotated and E'_R, E'_L are automatically located positions of facial landmarks. A successful localization was considered if $d_{eye} < 0.25$ which corresponds approximately to 1/4 of the annotated interocular distance $d(E_R,E_L)$ (a half of the width of the eye). After the locations of the landmarks are known, the location of the face in the image is also known.

4 Test Data

As the prospective application of the developed method lies in human-computer interaction, we assume that the input video includes some head rotations and facial expressions. For the purpose of method testing under these conditions, we created our own video database with neutral, frowning, and smiling faces under three controlled head rotations with angles of rotation 0°, 20°, and 30° in both right and left directions. We used a low-cost Canon Mini-DV camera with 720x568 pixel image resolution and 24-bit precision for color values. The potential impact of illumination, background, facial hair or eye-glasses was controlled to some extent in all video sequences and therefore ignored. No face alignment was performed.

The database consists of 10 Caucasian subjects (40% females) with average age of 30 years. Each video starts with neutral face, proceeds with facial expression, and ends up with neutral face. The level of the expression intensity varies among different subjects. In total, 150 video sequences were created with duration of about 7-8 seconds. The test data were annotated in advance by recording the true locations of the landmark centers in each frame for each test subject.

5 Results

The tests were run on the computer Dell Optiplex 745, Intel Core2 with 2133 MHz and 1 GB DDR2-memory in Win XP 2002 SP 2/Delphi-environment with DirectShow-interface. Fig. 4 shows the average rates of face localization under three expressions and three head rotations. The raw test data are shown in Table 2. The statistical analysis was done by using a two-way 3×3 (expression$\times$head rotation) repeated measures analysis of variance (ANOVA). The statistical analysis showed that head rotation had a statistically significant main effect on the face localization $F(2,18) = 9.31$, $p < 0.001$. Bonferroni-corrected pairwise post-hoc comparisons showed that the detection percentage was significantly lower when head was rotated by 30° $MD = 30.81$, $p < 0.05$ as compared with frontal head position. Difference between 20° and 30° head rotations was also statistically significant $MD = 22.14$, $p < 0.05$. There was no statistically significant difference between head rotation by 20° and frontal head position. ANOVA showed that there were no other significant main or interaction effects.

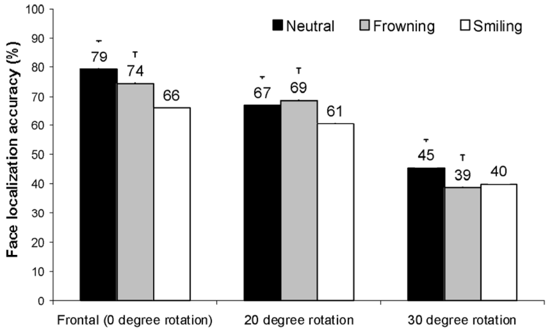


Figure 4. Average localization rates (%) of neutral, frowning, and smiling faces (all four landmarks were found) under three head rotations.

Table 2: Rates (%) of localization of neutral, frowning, and smiling faces under three head rotations.

Subject	Neutral			Frowning			Smiling		
	0°	20°	30°	0°	20°	30°	0°	20°	30°
1	69	91	97	79	84	67	96	72	93
2	95	51	17	82	74	31	87	67	17
3	55	17	30	45	10	37	27	34	31
4	98	68	60	98	83	35	87	72	48
5	95	62	36	89	91	17	82	56	29
6	33	47	49	21	51	47	26	62	46
7	93	88	30	85	79	13	24	49	46
8	63	61	25	81	35	8	90	44	14
9	92	95	98	62	93	71	85	77	37
10	91	72	30	97	82	64	63	65	45

6 Discussion

The developed method demonstrated high speed performance in face localization from streaming color video in real time. The speed of the method was 20 frames per second that meets the requirement of real-time video processing defined in [18]. This way, the method is comparable to the best existing real-time face detectors [6,15,19] in terms of processing speed.

The local oriented edges served as basic features for expression-invariant representation of facial landmarks. The results confirmed that in the majority of expressive images a distribution of the local oriented edges had structure with two horizontal dominants as predefined in [7]. This property allowed discarding noisy regions and preserving regions of the landmarks. Thus, the method was able to locate landmarks from images with hair and shoulders. The use of frontal-view geometrical face model further improved the overall performance of the method. As Fig. 4 shows, the method was effective in locating faces with frontal and near-to-frontal head poses. However, head rotations by 30° significantly decreased face localization rates. This is explained by the fact that geometrical constrains from Table 1 were defined mainly for frontal-view geometrical face model. The landmarks were located correctly in case of 30° head rotations, but failed to compose the face-like constellations. Relaxation of geometrical constrains from Table 1 or development of new measures and constrains for a near-to-profile geometrical face model would improve the performance of the method in case of strong face rotations.

In case of frontal and near-to-frontal head positions, the method demonstrated sufficiently high rates in locating faces with all three tested expressions - neutral, frowning, and smiling expressions. This gives similar performance of the method for these particular expressions as compared to the results of previous studies which use similar approach to facial landmark localization [7]. As distinct from that study, we concentrated on face localization rather than on independent landmark localization meaning that all four facial landmarks needed to be correctly located in order to declare successful face localization. Face localization rate would be improved if we consider two or three correctly located landmarks as necessary and sufficient requirement for successful face localization, as it is done, for example, in [2]. This way, the method can also be applied for independent facial landmark localization, when it is allowed to miss some landmarks.

In summary, as compared to the existing feature-based methods of face localization, the method demonstrated similar or superior performance in terms of localization rates [7,12] and speed [6,15,19]. Besides robustness to facial expressions and small out-of-plane head rotations, the developed method demonstrated robustness to noise such as hair, and elements of clothing and decoration. Emphasizing simplicity, high speed, and low computation cost of the method, we conclude that it can be used in face localization as such and also in preliminary localization of regions of facial landmarks for their subsequent processing where coarse landmark localization is followed by fine feature detection. The method is simple and straightforward to be utilized, for example, as face or facial landmark detector in the mobile phone environment.

7 Acknowledgement

This work was financially supported by the University of Tampere and Academy of Finland (project numbers 177857 and 1115997).

References:

[1] H. Chang and U. Robles, Face Detection, Project report, *http://www-cs-students.stanford.edu/robles/ee368/main.html*, 2000.

[2] D. Colbry, G. Stockman, and J. Anil, Detection of Anchor Points for 3D Face Verification, *CVPR'05*, Vol.3, pp. 118-126.

[3] T. Cootes, G. Edwards, and C. Taylor, Active Appearance Models, *Trans. Pattern Anal Mach. Intel.*, Vol.23, No.6, 2001, pp. 681-685.

[4] D. Cristinacce and T. Cootes, Feature Detection and Tracking with Constrained Local Models, *BMVC'06*, Vol.3, pp. 929-938.

[5] L. Farkas, *Anthropometry of the Head and Face*, (2 ed), Raven, New York, 1994.

[6] B. Fröba and C. Küblbeck, Robust Face Detection at Video Frame Rate Based on Edge Orientation Features, *FG'02*, pp. 342-347.

[7] Y. Gizatdinova and V. Surakka, Automatic Detection of Facial Landmarks from AU-Coded Expressive Facial Images, *ICIAP'07*, pp. 419-424.

[8] R. Gonzalez and R. Woods, *Digital Image Processing*, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.

[9] R. Gottumukkal and V. Asari, Real Time Face Detection from Color Video Stream Based on PCA Method, *AIPR'03*, pp. 146-150.

[10] E. Hjelmas and B. Low, Face Detection: A Survey*Comp. Vis. Image Understanding*, Vol.83, 2001, pp. 235-274.

[11] J. Martinkauppi, M. Soriano, and M. Laaksonen, Behavior of Skin Color under Varying Illumination Seen by Dierent Cameras at Dierent Color Spaces, *Mach. Vis. in Industrial Inspection*, V.9, No.4301, 2001, pp. 102-113.

[12] P. Michel and R. Kaliouby, Real Time Facial Expression Recognition in Video Using Support Vector Machines, *ICMI'03*, pp. 258-264.

[13] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz, Measuring the Performance of Face Localization Systems, *J. Image and Vis. Computing*, Vol.24, 2006, pp. 882-893.

[14] H. Rowley, S. Baluja, and T. Kanade, Neural Network-Based Face Detection, *Trans. Pattern Anal. Mach. Intel.*, Vol.20, No.1, 1998, pp. 23-38.

[15] S. Sclaroff and J. Isidoro, Active Blobs: Region-Based, Deformable Appearance Models, *Comp. Vis. Image Understanding*, Vol.89, No.2–3, 2003, pp. 197-225.

[16] D. Shaposhnikov, L. Podladchikova, and X. Gao, Classification of Images on the Basis of the Properties of Informative Regions, *Pattern Rec. Image Anal.*, Vol.13, No.2, 2003, pp. 349-352.

[17] K. Sobottka and I. Pitas, A Fully Automatic Approach to Facial Feature Detection and Tracking, *Lecture Notes In Comp. Science, AVBPA'97*, Vol.1206, pp. 77-84.

[18] M. Turk and M. Kölsch, Perceptual interfaces, *In G. Medioni and S.B. Kang, (Eds), Emerging Topics in Comp. Vis.*, Prentice Hall, chapter 9, 2004, 45 p.

[19] P. Viola and M. Jones. 2004. Robust Real-Time Face Detection, *Int. J. Comp. Vis.*, Vol.57, No.2, pp. 137–154.

[20] L. Wiskott, J. Fellous, N. Krüger, and C. der Malsburg, Face Recognition by Elastic Bunch Graph Matching, Trans. *Pattern Anal. Mach. Intel.*, Vol.19, No.7, 1997, pp. 775-779.

[21] M. Yang, N. Abuja, and D. Kriegman, Face detection using mixtures of linear subspaces, *FG'00*, pp. 70-76.

[22] M. Yang, D. Kriegman, and N. Ahuaja, Detecting Face in Images: A Survey, *Trans. Pattern Anal. Image Understanding*, Vol.24, 2002, pp. 34-58.

[23] D. Xi and S. Lee, Face Detection and Facial Component Extraction by Wavelet Decomposition and Support Vector Machines, *AVBPA'03*, pp. 199-207.

Publications in the **Dissertations in Interactive Technology** series

1. **Timo Partala:** Affective Information in Human-Computer Interaction
2. **Mika Käki:** Enhancing Web Search Result Access with Automatic Categorization
3. **Anne Aula:** Studying User Strategies and Characteristics for Developing Web Search Interfaces
4. **Aulikki Hyrskykari:** Eyes in Attentive Interfaces: Experiences from Creating iDict, a Gaze-Aware Reading Aid
5. **Johanna Höysniemi:** Desing and Evaluation of Physically Interactive Games
6. **Jaakko Hakulinen:** Software Tutoring in Speech User Interfaces
7. **Harri Siirtola:** Interactive Visualization of Multidimensional Data
8. **Erno Mäkinen:** Face Analysis Techniques for Human-Computer Interaction
9. **Oleg Špakov:** iComponent – Device-Independent Platform for Analyzing Eye Movement Data and Developing Eye-Based Applications
10. **Yulia Gizatdinova:** Automatic Detection of Face and Facial Features from Images of Neutral and Expressive Faces