



Graz University of Technology
Institute for Computer Graphics and Vision

Dissertation

SHAPE AND APPEARANCE BASED ANALYSIS
OF FACIAL IMAGES FOR ASSESSING ICAO
COMPLIANCE

Markus Storer

Graz, Austria, December 2010

Thesis supervisors

Prof. Dr. Horst Bischof

Prof. Dr. Fernando De la Torre

Facial analysis is not an exact science, but rather subjective, and open to many factors in a person's life, from genetics to environmental influences.

S. M. D'Souza

Abstract

The analysis of facial images has always attracted researchers across different disciplines ranging from psychology over human behavior understanding to computer vision and pattern recognition. It is very challenging to “read” faces automatically due to biological factors like gender, race, facial hair or age, and extrinsic factors like occlusion, noise, lighting and pose variations.

One of the most popular applications for facial analysis is biometrics. In this context the International Civil Aviation Organization (ICAO) provides a number of specifications to prepare automated recognition from travel document photos. The goal of these specifications is to increase security in civil aviation on the basis of standardized biometric data. Due to this international standard, there is a high demand for automatically checking facial images to assist civil service employees in decision-making.

We propose a facial analysis system according to the ICAO standard and several novel algorithms motivated by the needs of our facial analysis system to solve issues addressing the huge variety of facial appearance. Furthermore our facial analysis should demonstrate robustness to occlusion, noise, lighting and pose variations.

We start with a procedure called tokenization to bring arbitrary input images, containing a face, to a standardized coordinate frame. Therefore we need a robust face and facial component detection. After the detection of the eyes and mouth components, they are assessed if they show an open or closed state. To be robust, we propose to fuse several classifiers to utilize the strengths of the single classifiers. The ICAO standard prohibits faces differing from the frontal pose. Therefore, we propose three algorithms determining the pose of the head. The last criterion we have to deal with is occlusion. We show how to detect occlusions in a facial image. Furthermore, we propose a robust Active Appearance Model fitting strategy, which is based on our novel Fast-Robust PCA approach. All the approaches are evaluated extensively on several publicly available databases and our own dedicated database. We compare our algorithms to state-of-the-art approaches in terms of speed and accuracy.

Kurzfassung

Die Analyse von Gesichtsbildern hat schon immer die Forscher verschiedenster Disziplinen inspiriert, angefangen von der Psychologie über das Verstehen menschlichen Verhaltens bis hin zu Computer Vision und Mustererkennung. Gesichter automatisiert zu “lesen” ist eine große Herausforderung aufgrund biologischer Faktoren wie Geschlecht, Rasse, Alter oder Gesichtsbehaarung, und äußerer Faktoren wie Verdeckungen, Rauschen, Beleuchtung und Variationen in der Pose.

Eine der populärsten Anwendungen für die Gesichts-Analyse ist die Biometrie. In diesem Zusammenhang stellt die International Civil Aviation Organization (ICAO) eine Spezifikation zur Verfügung, um Fotos in Reisedokumenten für die automatisierte Erkennung vorzubereiten. Das Ziel dieses Standards ist die Sicherheit in der Zivilluftfahrt auf der Grundlage von standardisierten biometrischen Daten zu erhöhen. Aufgrund dieses internationalen Standards gewinnt die automatische Überprüfung von Gesichtsbildern mehr und mehr an Bedeutung um Beschäftigte des öffentlichen Diensts bei der Entscheidungsfindung zu unterstützen.

Wir zeigen ein Gesichts-Analyse-System basierend auf den Vorgaben des ICAO-Standards und verschiedene neuartige Algorithmen, motiviert durch die Anforderungen unseres Systems in Bezug auf die Vielfalt von Gesichtern. Unsere Gesichts-Analyse sollte robust gegenüber Verdeckungen, Rauschen, Beleuchtungsänderungen und Pose Variationen sein.

Wir beginnen mit einem Verfahren namens Tokenisierung. Dabei werden beliebige Bilder, die ein Gesicht enthalten, in ein standardisiertes Koordinatensystem gebracht. Daher brauchen wir eine robuste Gesichts- und Gesichtskomponenten (Augen, Mund) Detektion. Nach der Detektion der Augen und Mund-Komponenten werden diese auf einen offenen oder geschlossenen Zustand beurteilt. Um bei dieser Beurteilung robust zu sein, fusionieren wir mehrere Klassifikatoren um die Stärken der einzelnen Klassifikatoren zu kombinieren. Gemäß des ICAO-Standards müssen die Gesichter eine frontale Pose

aufweisen. Daher zeigen wir drei Algorithmen zur Pose Bestimmung des Kopfes. Als letztes Kriterium behandeln wir die Verdeckungen auf Gesichtsbildern. Wir zeigen, wie wir Verdeckungen detektieren können. Weiters zeigen wir unsere robuste Active Appearance Model (AAM) Fitting Strategie, die auf unserem neuen Fast-Robust PCA Algorithmus basiert. Alle Algorithmen werden umfassend auf verschiedenen frei zugänglichen Datenbanken und auf unserer eigenen Datenbank ausgewertet. Wir vergleichen unsere Algorithmen zu State-of-the-Art Ansätzen in Bezug auf Geschwindigkeit und Genauigkeit.

Acknowledgments

I wish to express my sincere gratitude to my fellow researchers and colleagues at the Institute for Computer Graphics and Vision. Specifically, I want to thank my advisor Horst Bischof, who supported me and provided me the opportunity to do the PhD at his institute. Many thanks go to my colleague Martin Urschler. A lot of work in this thesis originated from a great collaboration with him. He was always open for discussions and it was a great experience for me to work with him. Thanks to Josef Birchbauer who was the link to the Siemens Biometrics Center and who also contributed his experience and gave valuable hints. I also wish to thank Fernando De la Torre, who was my advisor at my stay at the Carnegie Mellon University. He was very helpful and agreed to serve as the second thesis supervisor.

The thesis would not have been possible without the support by my family and my friends. Thank you!

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place

Date

Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort

Datum

Unterschrift

Contents

1	Introduction	1
1.1	ICAO Face Analysis	2
1.2	Contributions	6
1.3	Related Work	10
2	Tokenization	13
2.1	Robust Facial Component Detection for Face Alignment Applications . . .	14
2.1.1	Face Image Normalization System Description	16
2.1.1.1	Face Detection	17
2.1.1.2	Eyes and Mouth Component Detection	18
2.1.1.3	Face Symmetry Line Detection	18
2.1.1.4	Robust Probabilistic Face and Eye Voting	19
2.1.1.5	Eye Localization Refinement	21
2.1.1.6	Face Normalization Transformation	21
2.1.2	Experimental Results	22
2.1.3	Summary	26
2.2	Intensity-Based Congealing for Unsupervised Joint Image Alignment	27
2.2.1	Congealing	28
2.2.2	Experimental Results	30
2.2.2.1	Congealing Handwritten Digits	30
2.2.2.2	Congealing Facial Images	31
2.2.3	Summary	32
3	Eyes and Mouth Analysis	35
3.1	Face Analysis System Description	36
3.1.1	Eyes Open Analysis	37

3.1.1.1	Active Appearance Model	37
3.1.1.2	Iris Detection Approach	38
3.1.1.3	AdaBoost Classifier	38
3.1.1.4	EigenEye Model	39
3.1.2	Mouth Closed Analysis	39
3.1.2.1	Geometric Dark Blob Analysis	40
3.1.2.2	Boosting Classifier	40
3.1.2.3	EigenMouth Model	41
3.2	Classifier Fusion	41
3.3	Experimental Results	43
3.4	Summary	48
4	Analysis of the Deviation from Frontal Pose	49
4.1	Boosting Approach	51
4.1.1	Experimental Results	52
4.2	Head Pose Estimation by Non-Linear Regression and Manifold Embedding	55
4.2.1	Biased Manifold Embedding	55
4.2.2	Non-Linear Regression of Image Descriptors	57
4.2.3	Facial Image Database	58
4.2.4	Experimental Results	59
4.2.5	Summary	61
4.3	3D-MAM	62
4.3.1	3D Morphable Appearance Model	64
4.3.1.1	3D Model	64
4.3.1.2	Model Fitting	66
4.3.2	Experimental Results	68
4.3.3	Summary	72
5	Occlusion Handling	73
5.1	Occlusion Detection	74
5.1.1	Methodology	76
5.1.1.1	Method 1	76
5.1.1.2	Method 2	80
5.1.2	Summary	87
5.2	Active Appearance Model Fitting Under Occlusion	89
5.2.1	Robust AAM Fitting	91
5.2.2	Experimental Results	92
5.2.3	Summary	94
5.3	Fast-Robust PCA	95
5.3.1	Preface	96
5.3.2	Derivation of the Algorithm	97

5.3.2.1	FR-PCA Training	98
5.3.2.2	FR-PCA Reconstruction	99
5.3.3	Experimental Results	102
5.3.4	Summary	106
6	Conclusion	109
6.1	Future Work	110
A	Acronyms	113
B	Active Appearance Model	115
C	Boosting	117
D	Annotation Tool	119
E	List of Publications	123
E.1	Book Chapter	123
E.2	Peer-Reviewed Conference Papers	123
	Bibliography	125

List of Figures

1.1	ICAO criteria (Illustration provided by Siemens Austria)	4
1.2	Siemens homeland security suite	5
1.3	Siemens homeland security suite (SHSS). Verification of biometric and personal data at border control stations. (Illustration provided by Siemens Austria)	5
1.4	Illustration of the overall image analysis system consisting of normalization (<i>tokenization</i>) and compliance analysis.	6
1.5	Our ICAO checked criteria.	7
1.6	Sample images: (a) frontal pose, eyes open, mouth closed, (b) left pose, eyes open, mouth closed, (c) right pose, eyes open, mouth closed, (d) frontal pose, eyes closed, mouth closed, (e) frontal pose, eyes open, mouth open, (e) frontal pose, eyes open, mouth open.	8
1.7	The organization of this thesis.	9
2.1	Illustration of the geometrical alignment procedure to form a normalized face portrait image (right) given an arbitrary input image (left). Image taken from the Caltech Faces database [19].	14
2.2	Illustration of the face image normalization work flow consisting of several stages. At the heart of the system is the voting scheme that decides from a number of face and facial component detections the most probable one while taking the detected symmetry line into account.	17

2.3	Probabilistic face and eye voting: (a) Setup of the voting procedure showing potential face rectangles (white and yellow, white ones have no support by detected facial components) and most likely face (red) due to support of facial components (green, cyan), and (b) prior probability distribution of spatial facial component locations in a face rectangle, modeled as normal distributions with means and covariance matrices derived from manually annotated training data. Image from Caltech Faces database [19].	19
2.4	Detail illustration of the eye region. (a) Active Appearance Model: Learned mean shape/texture and the texture after successful fitting. (b) AAM shape model after successful fitting drawn on the input image. (c) Eye setup for the experimental evaluation measure <i>face detection rate</i>	22
2.5	Quantitative results on relative error distributions of our method compared to different algorithms. (a) BioID database (1521 images), (b) AR database (509 images), (c) IMM database (240 images).	24
2.6	Selected qualitative results on representative images from the databases. . .	25
2.7	Average images before (first row) and after congealing (second row). The samples were obtained from the MNIST database [73].	31
2.8	Annotation of a facial image with 19 landmark points at salient facial feature positions.	31
2.9	The spatial variation gets removed from the perturbed facial samples by our algorithm. The samples are taken from the CVL database [109]. . . .	32
2.10	Point-to-Point distance of the unaligned landmarks and the congealed landmarks to the aligned landmarks.	33
3.1	Some sample images from the evaluation database: (a) eyes open, mouth closed, (b) eyes open, mouth open, (c) eyes closed, mouth open. Taken from the AR face database [90].	36
3.2	Face analyzer workflow. From tokenized images we perform some pre-processing, apply the single classifiers and fuse their results to form a final decision.	37
3.3	Sample images from the AR face database [90] showing the difficulties of the images under consideration.	43
3.4	Evaluation procedure for the eyes-open AAM on our own database. (a) Distribution of the confidence values for open-eyes in the database. (b) Distribution of the confidence values for closed-eyes in the database. (c) Characteristic of the false rejection rate (FRR) and false acceptance rate (FAR). (d) ROC curve	44
3.5	Comparison of the ROC curves of the best single classifier to the best fusion strategy for the (a) eyes-open and (b) mouth-closed analysis on the AR face database.	45

3.6	Comparison of the ROC curves of the best single classifier to the best fusion strategy for the (a) eyes-open and (b) mouth-closed analysis on our own face database.	45
4.1	The three degrees of freedom of a human head described by <i>pitch</i> , <i>roll</i> and <i>yaw</i> angles.	50
4.2	Pose classification with six AdaBoost learned pose classifiers.	52
4.3	Cropping area for the training images: centered crop (left column), shifted cropping area (right column)	53
4.4	The original image and its mirrored image were overlaid and the average pixel value was taken to form a new image (left figure: frontal image, right figure: image showing pose deviation). The red squares illustrate the cropping area for AdaBoost training.	54
4.5	Comparison of the relative detection rates for the analysis of the deviation from frontal pose.	54
4.6	Components of a head pose estimation system using manifold embedding .	56
4.7	Components of a descriptor based head pose estimation system	57
4.8	Examples of generated head poses	58
4.9	Yaw angle estimation performance of the HOG descriptor with SVR learning for different yaw angles (mean and standard deviation)	60
4.10	Evaluation of the frontal-pose classification performance of the three system using ROC plots	61
4.11	3D Morphable Appearance Model. Effect of varying the shape, texture and appearance parameters of the first and second mode by ± 3 standard deviations.	64
4.12	Patch extracted from the whole head used for synthesizing a new input image.	67
4.13	Fitting workflow	68
4.14	Mean and standard deviation of the absolute yaw angular error for the (a) USF Human ID 3D face database and (b) FacePix database.	69
4.15	Mean absolute error for pitch- and yaw angle on the USF Human ID 3D face database by altering pitch- and yaw angles for the generation of the test images. (a,b) Our approach (c,d) 3DMM	71
4.16	Analysis-by-Synthesis fitting example. The lower row shows the adjustment of the 3D model. In the upper row, the corresponding fitting patch is illustrated. This model fitting converges after 17 iterations.	72
5.1	Method 1. The tokenized and color adjusted image is transformed to the HSV color space. After binarization of the H-channel of the HSV color space, occlusion masks are used to calculate the level of occlusion on the lower facial part and around the eyes region.	77

5.2	Creation of an occlusion map based on the HSV color space. (first column) original images, (second column) H channel of HSV color space and (third column) the corresponding maps gained after thresholding the H channel image.	78
5.3	Samples from the AR database. The images feature frontal view faces with (a)-(b) different facial expressions, (c) several illumination conditions, (d) occlusion of the lower facial part by a scarf, (e) occlusion of the eyes by sunglasses and (f) combinations of these variations.	79
5.4	ROC curve of Method 1 evaluated on the AR database.	80
5.5	Samples from our own database. In addition to the variations exhibited by the AR database, our own database shows some more variations, e.g. (a) occlusions by skin-similar color of the lower facial part, (b) occlusions of the forehead, (c) variation of the color tone of the overall image, (d) extreme lighting conditions, (e) tinted glasses in several colors and (f) several colored occlusions of the lower facial part (also skin-similar color).	81
5.6	Method 2. The left branch of the whole approach is based on color techniques. If there is no occlusion found by this color occlusion detection, the ASM + PCA approach will be activated.	82
5.7	Determining similar colors. (a) The H-channel pixel values marked by the red lines are used to construct a Gaussian model. (b) Probability map of similar colors, (c) probability map after binarization and some morphological operations.	83
5.8	Comparison of the fitting quality of model based fitting. (first row) AAM, (second row) STASM.	84
5.9	ASM + PCA for occlusion detection.	86
5.10	ROC curve of Method 1 and Method 2 evaluated on our own dataset. . . .	87
5.11	Typical failure cases resulting from (a)-(c) skin-similar occlusion of the forehead, (d) larger deviations from frontal pose, (e) extreme facial hair and (f) slight specularities exhibited on the glasses.	88
5.12	Robust AAM fitting chain	92
5.13	Handling of occlusions for AAM fitting. (a) Test image. (b) Initialization of the AAM on the occluded image. (c) Direct AAM fit on occluded image. (d) AAM fit on reconstructed image. (e) Shape from (d) overlaid on the test image. Image taken from Caltech Faces data set [19].	93
5.14	Examples of AAM fits on natural occlusions like tinted glasses or wearing a scarf. (First column) Test images with AAM initialization. (Second column) Direct AAM fit on the test images. (Third column) AAM fit utilizing the FR-PCA pre-processing. Images are taken from the AR face database [90].	94
5.15	Examples of AAM fits on natural occlusions like beards.	95

5.16	FR-PCA training: A global PCA subspace and a large number of smaller PCA sub-subspaces are estimated in parallel. Sub-subspaces are derived by randomly sub-sampling the input data.	98
5.17	Random sampling restricted to image slices (vertical, horizontal and quadrant).	99
5.18	Reconstruction pipeline	99
5.19	Refinement step. Solve an over-determined system of equations.	100
5.20	Data point selection process: (a) data points sampled by all sub-subspaces, (b) occluded image showing the remaining data points after applying the sub-subspace procedure, and (c) resulting data points after the iterative refinement process for the calculation of the PCA coefficients. This figure is best viewed in color.	101
5.21	Demonstration of the insensitivity of the robust PCA to noise (i.e., occlusions): (a) occluded image, (b) reconstruction using standard PCA, and (c) reconstruction using the FR-PCA.	101
5.22	Illustrative examples of ALOI database objects [46] used in the experiments.	102
5.23	Box-plots for different levels of occlusions for the RMS reconstruction-error per pixel. PCA without occlusion is shown in every plot for the comparison of the robust methods to the best feasible reconstruction result.	105
5.24	Box-plots for different levels of salt & pepper noise for the RMS reconstruction-error per pixel. PCA without occlusion is shown in every plot for the comparison of the robust methods to the best feasible reconstruction result.	106
C.1	Haar wavelet-like features [142]	118
C.2	Integral image representation [142]	118
D.1	Screenshot of our annotation tool. The visible facial feature points are annotated.	120
D.2	Screenshot of our annotation tool. In the right pane, the 3D facial avatar is illustrated after automatically calculating the pose of the head. This avatar can be overlaid on the original input image.	120
D.3	Annotation procedure. (a) Annotated landmarks, (b) pose estimation, (c) back-projected landmark points	121

List of Tables

3.1	Prior confidences of the single classifiers. (left) eyes-open classifiers, (right) mouth-closed classifiers	44
3.2	Evaluation results for the eyes open analysis on the AR face database . . .	46
3.3	Evaluation results for the eyes open analysis on our own face database . . .	46
3.4	Evaluation results for the mouth closed analysis on the AR face database .	47
3.5	Evaluation results for the mouth closed analysis our own face database . . .	47
4.1	Prior confidences for the classifier fusion of the AdaBoost based deviation from frontal pose decision.	53
4.2	Comparison of the Biased Manifold Embedding, LGO and HOG approach in terms of mean absolute pose angle estimation error when trained and tested on our database and additionally tested on the FacePix database [12]	60
4.3	Mean, standard deviation, median, upper- and lower quartile of the absolute yaw angular error by only altering the yaw angle for the (a) USF Human ID 3D face database and (b) FacePix database. The results are compared to the 3DMM [15].	69
4.4	Evaluations for the USF Human ID 3D face database. (a) Mean, standard deviation, median, upper- and lower quartile of the absolute yaw and pitch angular error by altering the yaw- and pitch angle. (b) Average runtime* per facial fit. The results are compared to the 3DMM [15].	70
5.1	Point-to-Point error. Comparing the direct fit of the AAM on the test image to the AAM fit utilizing the FR-PCA pre-processing (point errors are measured on 240x320 facial images).	93
5.2	Parameters for the FR-PCA (a) and the R-PCA (b) used for the experiments.	103

5.3	Comparison of the reconstruction errors of the standard PCA, the R-PCA and the FR-PCA for several levels of occlusion showing RMS reconstruction-error per pixel given by (a) mean and standard deviation and (b) median, lower- and upper quartile.	103
5.4	Comparison of the reconstruction errors of the standard PCA, the R-PCA and the FR-PCA for several levels of salt & pepper noise showing RMS reconstruction-error per pixel given by (a) mean and standard deviation and (b) median, lower- and upper quartile.	104
5.5	Runtime comparison. Compared to R-PCA, FR-PCA speeds-up the computation by a factor of 18.	105

Introduction

Facial analysis is a rather subjective discipline, which depends on many factors, from genetics to environmental influences. Therefore, it is very hard to find a basic recipe to interpret the expression and emotion of a person's face. Humans become especially skilled at analyzing personality characteristics based on facial features, hair, facial expressions and gaze. They learn to recognize these visual cues and to read faces within seconds. In this context an interesting quotation stands out: "*The face is the mirror of the mind.* - Saint Jerome". The face is a rich source of nonverbal communication. A famous psychologist and philosopher said: "*One cannot not communicate.* - Paul Watzlawick". That means, even when not talking, a person always conveys a message.

Automatically determining the facial expression and subsequently deriving the emotional state of a person attracted a lot of researchers in the computer vision and pattern recognition community. Facial analysis is not an exact science, but as we said before, very subjective. Facial expressions can also be interpreted differently depending on the cultural background. The facial analysis research is also driven by the increasing demands of many applications like:

- human-computer interaction
- human behavior understanding
- automotive safety
- facial avatar animation
- biomedical applications

- video surveillance
- biometrics
- computer games

Human-computer interaction [63] is providing an interface between a human and a computer, e.g., computers can be used to determine the gaze direction as well as head gesturing. This enables very direct means to interact with virtual worlds, especially in the computer gaming industry. A further example includes the interaction with robots. *Human behavior understanding* [107] is strongly related to human-computer interaction. Here the interactions are interpreted, e.g., the meaning of head movements like nodding, to interpret facial expressions or social interactions. In the domain of *automotive safety*, e.g., the driver's head is monitored to recognize driver distraction and inattention to avoid vehicle collisions [99]. In the *facial avatar animation* [39], a system records and extracts facial expressions from a source individual and transfers the expressions to another individual or artificial avatar. For example, this procedure is used in the movie industry to animate characters. *Biomedical facial analysis* [49] assesses medical states like fatigue or bad posture used in, e.g., driver assistance systems [135] or for the prevention of work-related disorders.

Biometric related applications like person verification/identification [1, 78, 160] or video surveillance also require the analysis of facial images. Our specific interest lies in security related applications to check facial images for their compliance to the International Civil Aviation Organization (ICAO) specification for machine readable travel documents [57]. This compliance check is our main motivation in this thesis. Thus, we propose several algorithms motivated by the needs of our face analysis system to solve issues addressing the huge variety of facial appearance, e.g., gender, race, facial hair or age. Furthermore our face analysis should demonstrate robustness to occlusion, noise, lighting and pose variations. All these aspects make it particularly difficult to design a robust system for face analysis.

We present a general description and the goal of the ICAO face analysis system in Section 1.1. An overview of our system and our developed algorithms is given in Section 1.2.

1.1 ICAO Face Analysis


The ICAO specification [57] describes a standardized coordinate frame based on face and eye positions for potential travel document photographs, therefore a face image alignment

(tokenization) step has to be performed. Furthermore, the standard provides a definition of parameters for image quality and facial expressions that classify images as suitable or improper for documents like passports, identification cards or visas. The ICAO criteria are presented in Figure 1.1. These ICAO criteria specify requirements to prepare automated face recognition from travel document photos and to allow interoperability among vendors. The goal of these specifications is to increase travel security on the basis of standardized biometric data. Furthermore, the overall system, called *ICAO portrait checker*, provides a method that assists civil service employees in determining suitable machine readable travel document photos, thereby increasing efficiency in this selection process and significantly reducing manual work. For example, the US Visa Waiver Program requires more secure and biometrically enabled passports. The aim there is to improve international travel security.

The advantages of this automatic checking system are:

- Objectively analyze and evaluate the quality of face images according to the ISO 19794-5 standard.
- No more subjective manual checking, wondering whether the captured images are suitable for automated facial recognition tasks.
- Higher throughput at inspections points.
- Interoperability and compliance with international standard and legislative requirements.
- Higher facial recognition performance.
- Higher database integrity.
- Easy to integrate into existing or new applications.

The ICAO portrait checker is a module in a homeland security suite, see Figure 1.2. This suite consists of several modules. The face module processes facial images originating from digital cameras, scanners or an input file and offers automatic facial image verification functionality. The fingerprint modules provide accurate fingerprint matching. The eDocument module reads data stored on a RF-ID of a machine readable travel document. All the modules can contribute to several applications. The data capture, validation and



Scene requirements	pose	deviation from frontal pose
	expression	closed eyes / covered eyes open mouth eyes looking away
	background	background
	lighting, shadows, hot spots	non uniform face lighting
	eye glasses	eye glasses present eye glasses lighting artifacts eye glasses tinted / colored rim of glass covering part of the eye eye glasses heavy frames (planned)
	head coverings	face covered
Photographic requirements	no over or under exposure	
	focus and depth of field	
	unnatural color	white balance red eyes
Digital requirements	color profile	greyscale density / color saturation

Figure 1.1: ICAO criteria (Illustration provided by Siemens Austria)

checkpoint applications offer a complete workflow for enrollment, border control and self-service checking stations for machine-readable travel documents. One possible application is illustrated in Figure 1.3.

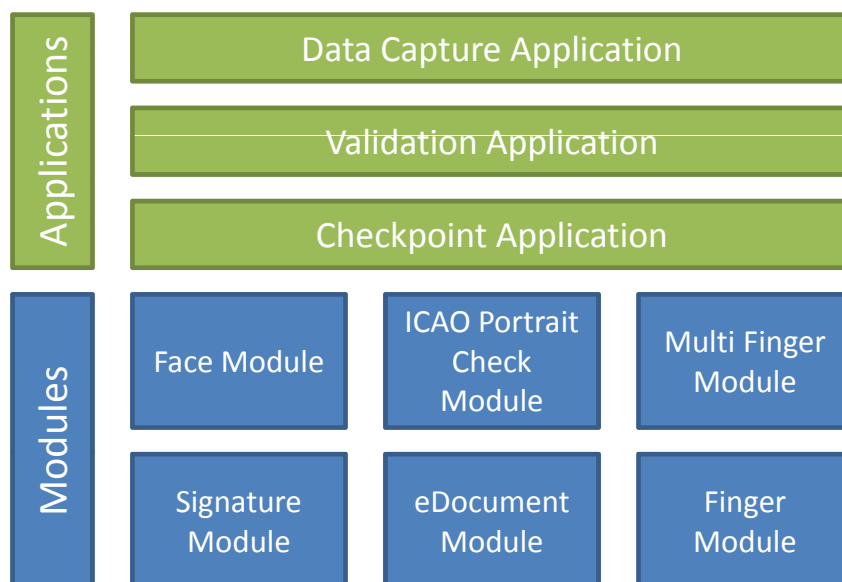


Figure 1.2: Siemens homeland security suite

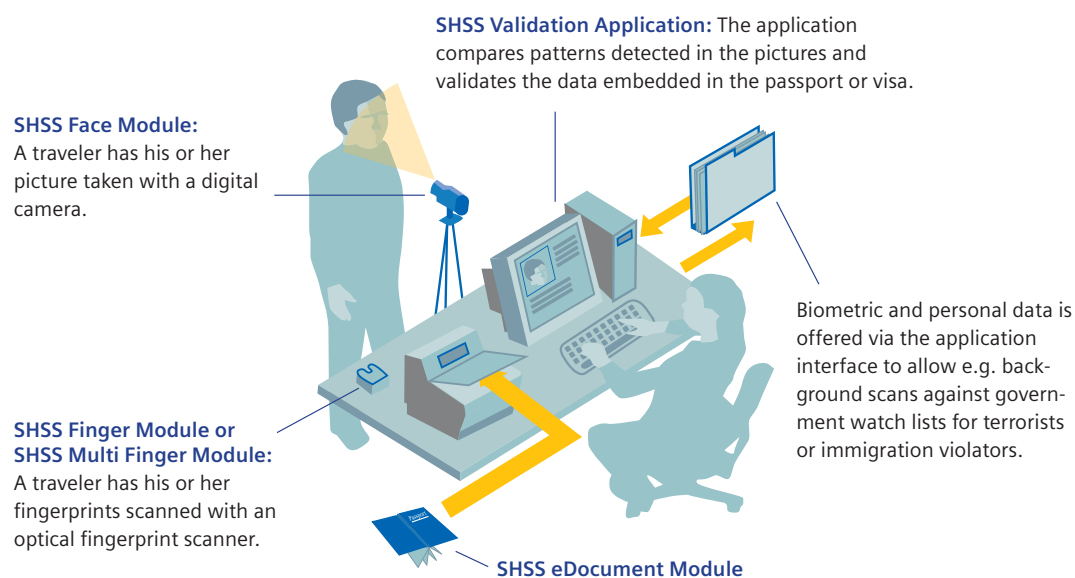


Figure 1.3: Siemens homeland security suite (SHSS). Verification of biometric and personal data at border control stations. (Illustration provided by Siemens Austria)

1.2 Contributions

We propose a facial analysis system according to the ICAO standard and several algorithms motivated by the needs of our facial analysis system to solve issues addressing the huge variety of facial appearance, e.g., gender, race, facial hair or age. Furthermore our face analysis should demonstrate robustness to occlusion, noise, lighting and pose variations.

Our facial analysis system consists of two modules. First, we need to derive a normalized face coordinate system based on eye locations to extract a standardized face image from arbitrary input images. We will refer to this process as *tokenization* according to [57]. The second step takes the tokenized images as input and performs several analysis algorithms in order to derive continuous scores for the likeliness of the corresponding event, e.g., *mouth-closed*. Figure 1.4 shows the overall workflow of the face analysis system.



Figure 1.4: Illustration of the overall image analysis system consisting of normalization (*tokenization*) and compliance analysis.

We focus on parts of the ICAO specification for checking arbitrary input images for their compliance: *eyes-open*, *mouth-closed*, *deviation from frontal pose* and *face covered* criteria, see Figure 1.5. The following rules for accepting photos are applicable. The full-face frontal pose has to be used. Rotation of the head has to be less than ± 5 degrees from frontal in every direction – up/down, rotated left/right, and tilted left/right. The face expression has to be neutral (non-smiling) with both eyes open normally, looking straight into the camera and the mouth closed. A smile is unacceptable regardless of the inside of the mouth and/or teeth being exposed or not. The region of the face, from the crown to the base of the chin, and from ear-to-ear, has to be clearly visible and free of shadows. Special care shall be taken in cases when veils, scarves or headdresses cannot be removed for religious reasons to ensure these coverings do not obscure any facial features and do not generate shadows. In all other cases head coverings must be absent. Some sample images exhibiting variations according to the specification are shown in Figure 1.6.

Biometric applications like face recognition/verification systems are not able to work with arbitrary input images taken under different imaging conditions or showing occlusions and/or variations in expression or pose. Thus, the main intention of the ICAO standard is to define how arbitrary images have to be prepared in order to perform robust and highly

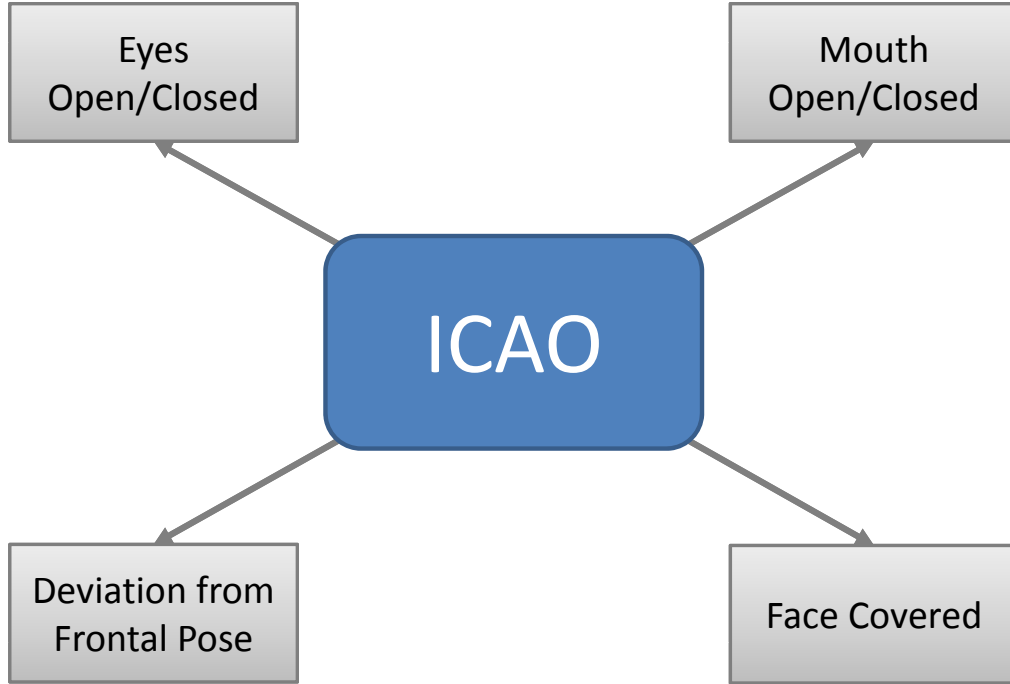


Figure 1.5: Our ICAO checked criteria.

accurate face recognition/verification. To support face recognition one has to perform a face image alignment (tokenization) step that takes occlusions/variations into account. Thus, we present a robust face normalization algorithm suitable for arbitrary input images containing a face in Section 2.1. The algorithm is based on detecting face and facial component candidates and robustly voting for the best face and eyes. For many of our used algorithms in the tokenization step, e.g., the AdaBoost learning algorithm, we need a lot of training data of high quality. Thus, we propose an approach for an unsupervised alignment of an ensemble of images called congealing in Section 2.2.

After the detection of eyes and mouth components in a facial image, an analysis procedure is applied to assess the ICAO criteria *eyes-open* and *mouth-closed*, see Chapter 3. We propose to fuse the decisions of several classifiers to gain robustness in the presence of difficult situations like noisy input data, lighting conditions or partial occlusions, e.g., wearing glasses.

According to the ICAO standard, persons are required to show a frontal head pose which means that the head rotation must not deviate more than ± 5 degrees in any direction from frontal. Thus, we propose three head pose estimation approaches starting from a coarse head pose estimation (only frontal/non-frontal decision) to a very fine estimation



Figure 1.6: Sample images: (a) frontal pose, eyes open, mouth closed, (b) left pose, eyes open, mouth closed, (c) right pose, eyes open, mouth closed, (d) frontal pose, eyes closed, mouth closed, (e) frontal pose, eyes open, mouth open, (f) frontal pose, eyes open, mouth open.

in Chapter 4. The coarsest approach is based on AdaBoost classification. The next finer estimation approach builds upon a Histogram of Oriented Gradients (HOG) descriptor used as input for a non-linear regression. Furthermore a Biased Manifold Embedding (BME) approach is extended to cope with multiple pose-angles. In addition, we present an approach for the creation of an artificial training database. The finest head pose estimation is a novel approach called 3D Morphable Appearance Model (3D-MAM) and is applied for head pose estimation. The finer head pose estimation approaches are more accurate but are only applicable in a smaller range of head poses.

Photographs showing occlusions are prohibited according to the ICAO standard. In Chapter 5 we show our approaches dealing with occlusions. Section 5.1 shows an ap-

proach for occlusion detection. We detect occluded images automatically and evaluate the occlusion detection performance on several databases. In case of an occluded image, we cannot only detect occlusions, but we can also improve the fitting quality of an Active Appearance Model (AAM), see Section 5.2. The AAM itself is not robust against occlusions. If parts of the image are occluded, the AAM fails to converge correctly and the obtained results become unreliable. To overcome this problem we propose a robust AAM fitting strategy. The main idea is to apply a robust PCA model to reconstruct the missing feature information and to use the obtained image as input for the standard AAM fitting process. Since existing methods for robust PCA reconstruction are computationally too expensive for real-time processing we developed a more efficient method: Fast-Robust PCA (FR-PCA), see Section 5.3. In fact, by using our FR-PCA the computational effort is drastically reduced. Moreover, more accurate reconstructions are obtained. In the experiments, we evaluate both, the FR-PCA model and the whole robust AAM fitting chain on facial images. The results clearly show the benefits of our approach in terms of accuracy and speed when processing disturbed data (i.e., images containing occlusions).

The thesis is concluded in Chapter 6 and possible directions for future work are given. The organization of this thesis is given in Figure 1.7.

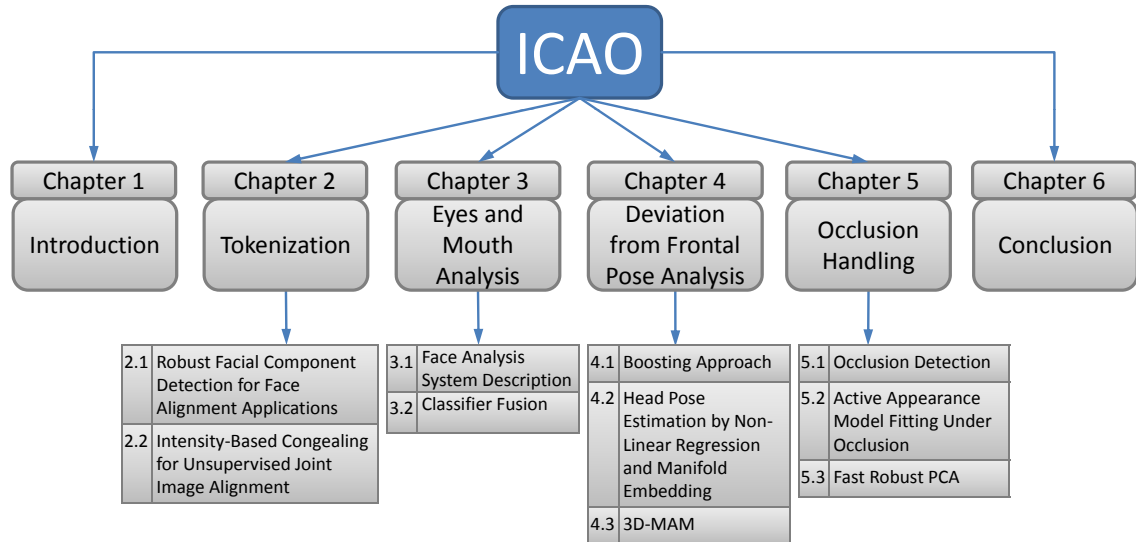


Figure 1.7: The organization of this thesis.

1.3 Related Work

Facial analysis has a long tradition in the computer vision literature. We give an overview over the facial analysis literature in this related work section. More extensive surveys of the appropriate related literature will be given in the following dedicated chapters and sections for our algorithms and our applications.

Face and facial component detection covers a broad area in the computer vision literature, a survey can be found in [51, 153]. Face detection from single images is challenging due to possible variabilities in scale, location, orientation, and pose. Schneiderman and Kanade [121] use multi-resolution information based on wavelet analysis to construct a nonlinear face and non-face classifier based on histogram statistics and the AdaBoost learning technique. Their approach inspired the important work of Viola and Jones [142] using simple and efficient Haar features which had an enormous impact on object detection over the last years. An example for a non-holistic approach is Heisele et al. [50] who exploit the idea of using components like mouth, eyes and nose for face detection. A similar direction is proposed by Felzenszwalb and Huttenlocher [42] in their work on pictorial structures. A recent work in this direction is Erubimov and Lee [40] who combine facial components with a graphical model. Recent methods can also be applied successfully for multi-view face detection [52, 148]. These methods are able to find faces exhibiting arbitrary poses.

Facial expression analysis is used to analyze facial expressions resulting from one or more motions or positions of the muscles of the face. Facial expressions are a form of nonverbal communication. Areas of applications include the recognition of the behavior and emotion from images and videos. It can also be used to animate a facial avatar. Surveys can be found in [41, 108]. According to [38], there are six basic expressions: happiness, sadness, fear, disgust, surprise, and anger. The facial expressions are coded by Action Units (AUs) in the Facial Action Coding System (FACS) [38], which objectively describe the facial deformations in terms of visually observable muscle actions. Most of the facial expression analysis systems extract 2D spatiotemporal facial features either by landmark locations (e.g., corners of the eyes or mouth) [17], by the shape of the facial components (e.g., eyes, mouth) [68] or the texture modality [154].

Face recognition has received significant research attention over the last two decades [2, 8, 137, 158]. Zhao et al. provides a survey [160]. This area has also led to many commercial systems [110]. The area of face recognition can be divided into verification and identification [59]. In verification, a system validates whether the

individual is the one who claims to be. In contrast, in face identification the individual is searched in a database to expose the identity.

Head pose estimation An extensive survey on head pose estimation problems can be found in [100]. *Appearance template methods* [10, 102] try to estimate the head pose by directly comparing facial images with either a set of template images or a detector array in order to assign a new facial image to a pose with the most similar feature. *Classification based methods* [54, 62] learn classifiers between the input image and a discretized space of poses. There are many classifiers that have been used such as multiclass SVMs [76], multiclass linear discriminant analysis (LDA) or the kernelized version (KLDA [34]), such that a test facial image can be assigned to one of the discrete pose classes [150]. These methods also suffer from the non-uniform sampling of the input space and only return discrete head pose estimates. *Regression based methods* estimate the pose by learning a linear or nonlinear function between the image and continuous angles. Several regressors are possible such as Support Vector Regression (SVR) [79, 87, 96] and Gaussian Process Regression (GPR) [113]. *Embedding based methods* compute the low dimensional counterparts of the high dimensional training data lying on a continuous manifold describing the pose variations of facial images [92]. A test image is first embedded into these low dimensional manifolds and then used for template matching or regression to compute the angle. Balasubramanian et al. [6] presents a supervised extension to Isomap [130], locally linear embedding (LLE) [117], and Laplacian Eigenmaps (LE) [9] and shows, that this change improves the head pose estimation performance.

Regarding the ICAO facial analysis system itself, there have been no previous publications on this topic besides [128], reporting results on an automatic face image validation system, where a number of rather simple quality aspects of face images are checked. However, a number of commercial products (e.g., Cognitec^{TM*}, Kee Square^{TM†}) currently exist for ICAO compliant facial analysis.

*www.cognitec-systems.de

†www.keesquare.com

Tokenization

Biometric applications like face recognition/verification are an important topic in computer vision, both for research and commercial systems. Unfortunately, state of the art face recognition systems are not able to work with arbitrary input images taken under different imaging conditions or showing occlusions and/or variations in expression or pose. To support face recognition one has to perform a face image alignment (tokenization) step that takes occlusions/variations into account. In Section 2.1 we present a robust face normalization algorithm suitable for arbitrary input images containing a face. The algorithm is based on detecting face and facial component candidates and robustly voting for the best face and eyes. Our restrictions are a certain pose range (frontal to half profile) and suitable illumination conditions.

Faces are detected using the well known AdaBoost technique [141] incorporating Haar-like- and orientation histogram features. There might be more than one detection from the AdaBoost face detection algorithm [56], so the correct face has to be found in a robust way. In addition, for every potential face detection rectangle, the facial components, mouth and eyes, are located using AdaBoost classifiers [142]. The AdaBoost learning algorithm is a supervised method, i.e., we need a lot of labeled training data. We manually labeled the facial images and aligned them in an unsupervised manner. This alignment of an ensemble of images is called congealing and is described in Section 2.2.

The most likely face is obtained using the face and facial component detections in a normally distributed probabilistic model based on an empirically determined prior distribution of eye and mouth locations. The derived eye locations from the AdaBoost step are further refined by an Active Appearance Model (AAM) [22]. Identified left and right eye coordinates initiate the actual tokenization step according to the ICAO specification [57].

Our algorithm is designed to deal with occlusions and its performance is shown on three publicly available image databases. Figure 2.1 illustrates the tokenization step.

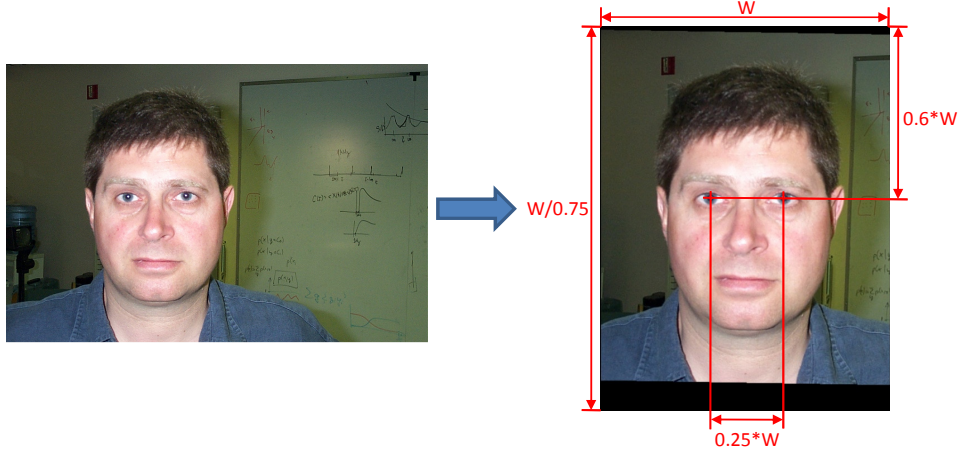


Figure 2.1: Illustration of the geometrical alignment procedure to form a normalized face portrait image (right) given an arbitrary input image (left). Image taken from the Caltech Faces database [19].

2.1 Robust Facial Component Detection for Face Alignment Applications

The main intention of the ICAO standard is to define how arbitrary images have to be prepared in order to perform robust and highly accurate face recognition/verification. For the purpose of geometrical alignment a part of the ICAO specification describes a standardized coordinate frame based on eye locations. A definition of this geometrical coordinate system forming normalized face portrait images (token images) is shown in Figure 2.1. This illustration clearly motivates the need to localize faces and facial components like eyes or mouth in arbitrary input images. Localization algorithms suffer from problems due to occlusions like glasses, beards or covering objects (hands or hair) or facial expressions like an open mouth, closed eyes or facial grimaces. Images with differences in facial pose and illumination conditions further contribute to the difficulty of facial image analysis and normalization.

In this section we describe a robust facial image detection and normalization algorithm that may serve as a powerful pre-processing step for face recognition [160], face analysis and facial expression recognition [106]. Our method is based on separate modules for

facial component detection which are combined in a probabilistic voting scheme to fuse the results of independent detectors, thus achieving robustness to occlusions and disturbances. The voting scheme results in face and eye locations that are used to calculate a face normalization transformation.

Detection of faces and facial components has a long tradition in computer vision research. Application areas dealing with face analysis/processing require an initial face localization step, see [51, 153] for literature surveys on face detection methods. Face detection from single images is challenging due to possible variabilities in scale, location, orientation, and pose. Further, different facial expressions, occlusions and illumination conditions also have an effect on the appearance of faces. According to [153], we define face detection as: Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, return the image location and extent of each face. Further, facial feature detection has the goal of detecting facial components like eyes, mouth, nose, ears, etc., either under the assumption that there is only a single face in the image or given a known face detection.

In the face detection literature one distinguishes holistic (whole face is detected at once) and component based (non-holistic) approaches. Moghaddam and Pentland [95] describe a holistic principal component analysis (PCA) based detection system, where a PCA subspace (or eigenface representation) is used and the detection is performed in a sliding window manner. Rowley et al. [119] show a neural network based approach for face detection, where the sliding windows are pre-processed and reduced in their dimensionality in order to be put into a neural network. Jesorsky et al. [60] describe a face detection method based on point sets and the Hausdorff distance. Schneiderman and Kanade [121] use multi-resolution information based on wavelet analysis to construct a nonlinear face and non-face classifier based on histogram statistics and the AdaBoost learning technique. Their approach inspired the important work of Viola and Jones [142] using simple and efficient Haar features which had an enormous impact on object detection over the last years. An example for a non-holistic approach is Heisele et al. [50] who exploit the idea of using components like mouth, eyes and nose for face detection. These components are related by constraints on their spatial configuration, which obviously makes the algorithm more robust to occlusions. A similar direction is proposed by Felzenszwalb and Huttenlocher [42] in their work on pictorial structures. A recent work in this direction is Erukhimov and Lee [40] who combine facial components with a graphical model. Our presented approach also uses the strategy to combine facial components for face detection

and eye localization to be more robust against occlusions. A slightly different research direction for face detection makes use of statistical models of shape and/or appearance to localize faces and facial components, with the seminal work by Cootes et al. [22].

In the following paragraphs we describe two important basic components of our system in more detail. These are the efficient object detection approach from Viola and Jones [142] and the statistical shape and appearance model of Cootes et al. [22].

Efficient AdaBoost Object Detection The most influential work in object detection of the last few years is definitely the approach of Viola and Jones [142] based on boosting simple weak classifiers to form a strong classifier (Appendix C). One of their application areas was face detection.

Localization Refinement based on Statistical Active Appearance Models Generative model-based segmentation approaches for highly accurate feature localization have received a lot of attention in computer vision over the last decade. While the Viola and Jones approach solely leads to a coarse feature localization due to the sliding window approach, an exact object delineation is often performed using statistical models of shape and appearance. The most important representative of this class of algorithms is the Active Appearance Model (AAM) from Cootes et al. [22], see Appendix B. It has successfully found a large number of applications ranging from face detection to medical image analysis.

The remainder of this section is structured as follows. Section 2.1.1 describes the components of our algorithm and the overall system. To demonstrate the strength of our approach we evaluate our algorithm on publicly available databases and compare it to state of the art methods in Section 2.1.2. Finally, we discuss our findings and summarize our work in Section 2.1.3.

2.1.1 Face Image Normalization System Description

In this section we describe our novel algorithm for robust face image normalization. We consider arbitrary input images containing at least one face. However, there are some assumptions we have to make considering potential input images. We restrict the possible deviations in pose angles of the depicted head to smaller than 45 degrees in yaw and pitch. For heads with larger angles the face and facial component detection performance significantly deteriorates, however, in our targeted application of normalizing face images according to the ICAO specification we can safely assume that profile or near-profile

images are rare. Roll angle deviations in head pose are targeted in our system by the facial component detection and the face symmetry lines. Another assumption we have to make is that images are taken under normal lighting conditions. Illumination problems that lead to nonlinear intensity changes are therefore not considered.

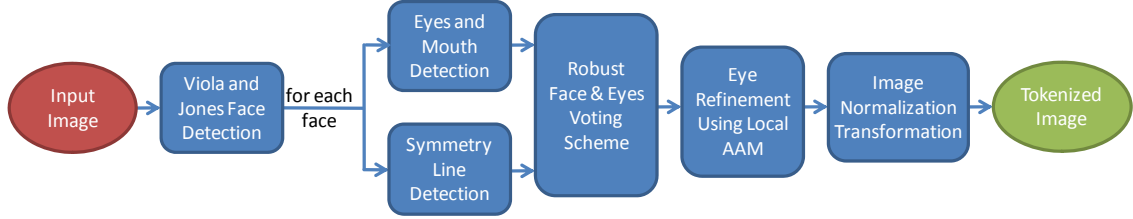


Figure 2.2: Illustration of the face image normalization work flow consisting of several stages. At the heart of the system is the voting scheme that decides from a number of face and facial component detections the most probable one while taking the detected symmetry line into account.

Our algorithm consists of several modules including face detection, facial component detection, face symmetry line detection, incorporation of spatial a priori knowledge in a probabilistic voting scheme for most likely face and eye detections, a detailed eye center localization module using a statistical model describing shape and appearance, and finally a normalization transformation. The algorithm work-flow is illustrated in Figure 2.2.

2.1.1.1 Face Detection

The face detection component uses the efficient face detector [142] described in Section 2.1. We extend the original feature set according to Lienhart et al. [80] using an open-source implementation [56] that provides a pre-trained detector cascade for portrait and near-portrait faces. After experimentation with this pre-trained face detection cascade we were able to find a parameter set leading to a high detection rate, however, also leaving a significant number of false positive detections. Thus, to achieve an accurate face detection we additionally need to detect facial components (eyes and mouth) to vote for the single face candidate having the highest probability of being the correct face. To prepare for the subsequent probabilistic voting scheme we calculate a normalized confidence measure

$$f = 1 - \exp \left(\frac{\sum_m \alpha_m h_m(\mathbf{x})}{\max_m(\alpha_m)} \right),$$

where \mathbf{x} is a pattern to be classified, $h_m(\mathbf{x}) \in \{-1, +1\}$ are the M easily constructible, weak classifiers and $\alpha_m \geq 0$ are the combining coefficients. More details on the Boosting algorithm can be found in Appendix C. This measure resembles the confidence of a certain spatial location to be a face.

2.1.1.2 Eyes and Mouth Component Detection

We trained facial component detectors using the same AdaBoost object detection (see Section 2.1) framework that was used for face detections. However, the facial components were trained by using our own set of positive and negative object patches. We used 5646 manually annotated face images which were taken from several databases; the Caltech Faces database [19], the FERET database [111], and our own proprietary database. Manual annotations were performed for left and right eye patches and for mouth patches under different poses and different facial expressions (eyes closed, half-closed and open, mouths closed, open, wide open or smiling). Roughly 40 percent of these images show a deviation from frontal pose and neutral expression. 16 percent of the images contain glasses and 21 percent contain a beard.

For facial component detection we apply these individual detectors to each potential face detection and look for the eyes in the upper two thirds of the face detection, while the mouth search area is restricted to the lower two thirds. This way we receive a number of potential eye and mouth candidates for each face. Again we tune the detection parameters to get a high detection rate with the downside of a larger number of false positive detections, an issue that gets resolved later in our voting scheme. To calculate a normalized confidence measure we calculate the same confidence as described in Section 2.1.1.1.

2.1.1.3 Face Symmetry Line Detection

Facial mirror symmetry is an important cue that should contribute to a robust face detection system to deal with in-plane rotations. After investigating the state of the art in symmetry detection algorithms we developed a simple and efficient face symmetry detection scheme based on [112]. In this algorithm we calculate image gradient points restricted to the detected face patch. Afterwards we randomly draw a large number of image gradient point pairs, where each of these point pairs votes for a certain symmetry line normal to their connecting line segment. The voting is performed similar to a Hough line voting scheme, taking the similarity of the gradient magnitudes of the point pair into account as the voting increment. Finally the line with the highest vote counter is selected as the

potential symmetry line and the magnitude of the vote counts is used as a confidence value. Since this algorithm is rather time-consuming when implemented testing all point pairs, we restrict our search range to point pairs leading to symmetry lines between minus and plus 25 degrees.

2.1.1.4 Robust Probabilistic Face and Eye Voting

The previous detection stages lead to a number of face and facial component hypotheses. We proceed with a probabilistic voting framework to find the most likely candidates. Our goal is to effectively combine the different components to form a robust vote for a face detection and an associated eye localization under the constraints of a priori knowledge about spatial locations.

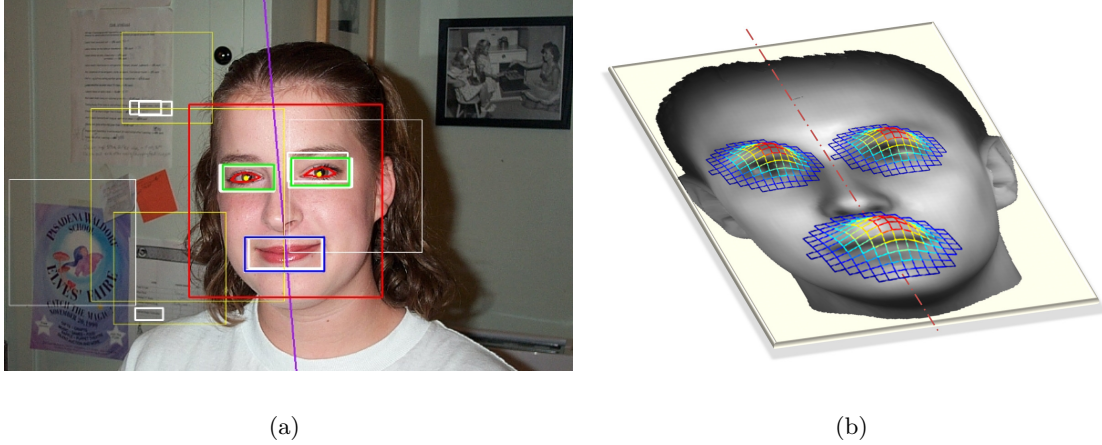


Figure 2.3: Probabilistic face and eye voting: (a) Setup of the voting procedure showing potential face rectangles (white and yellow, white ones have no support by detected facial components) and most likely face (red) due to support of facial components (green, cyan), and (b) prior probability distribution of spatial facial component locations in a face rectangle, modeled as normal distributions with means and covariance matrices derived from manually annotated training data. Image from Caltech Faces database [19].

We denote the K potential face rectangle candidates in an image as $F_i, i = 1, \dots, K$. For each of the K faces, we have candidate rectangles for left eye $L_{j_l}^i, j_l = 1, \dots, N_l$, right eye $R_{j_r}^i, j_r = 1, \dots, N_r$ and mouth $M_{j_m}^i, j_m = 1, \dots, N_m$. Since we have no a priori knowledge about the spatial location of faces in the arbitrary input images, we assume it uniformly distributed. The confidence in a certain face rectangle candidate now depends on two factors. A measure $c_f(F_i)$ delivered from the AdaBoost detection and the contribution of the facial components that were detected inside the face rectangle. For the

individual facial component detections per face rectangle we perform a Bayesian probability estimation to reach a maximum a posteriori (MAP) solution. Therefore, a priori assumptions $p(L), p(R), p(M)$ on the spatial distribution of eye and mouth patches relative to face rectangles (in a normalized coordinate system to be independent of the size of rectangles) are estimated in a maximum likelihood sense from our database of annotated portrait and near-portrait face images. Thus, the models are normal distributions calculated from the facial component locations in the face coordinate system (see Figure 2.3b for an illustration). Here the detected symmetry line (compare Section 2.1.1.3) is taken into account in order to perform an in-plane rotation of the prior distributions. Furthermore, each facial component detection has an associated likelihood measure from the AdaBoost detection, e.g. for the left eye $p(L_{j_l}|L)$ resembling the likelihood of measurement L_{j_l} being a left eye. To sort out the most likely facial components we apply Bayes theorem to get an MAP solution by comparing the products of priors and likelihoods. For each face F_i we thus receive a left eye rectangle $L_{j_l}^i$ with

$$\arg \max_{L_{j_l}^i} (p(L|L_{j_l}^i)) = \arg \max_{L_{j_l}^i} (p(L_{j_l}^i|L)p(L)),$$

and right eye $R_{j_r}^i$ and mouth $M_{j_m}^i$ rectangles, respectively. In case no candidate can be found for a facial component from the AdaBoost detection, we hypothesize a candidate at the location of the maximum of the prior distribution. However, this hypothetical candidate gets assigned only a small likelihood of 0.1 since we do not want the guessed part to have a high confidence. If in a face candidate rectangle no facial component can be found at all, we remove the face candidate from further processing. The posteriors for the three facial components now get combined to a single confidence vote by averaging the (normalized) posteriors

$$c_{fc}(F_i) = \frac{1}{3}(p(L|L_{j_l}^i) + p(R|R_{j_r}^i) + p(M|M_{j_m}^i)).$$

If no facial component can be found in a face candidate rectangle, we set $c_{fc}(F_i) = 0$. We finally combine the confidences of face rectangle and facial component confidence votes to a single vote using $c(F_i) = c_f(F_i)c_{fc}(F_i)$. The face rectangle F_i maximizing $c(F_i)$ is taken as the final result of the face voting, with the corresponding left and right eyes $L_{j_l}^i, R_{j_r}^i$ as final results for the eye localization. This simple scheme gives us robustness to occlusions, since it is not necessary to detect all facial components as long as the final face confidence measure of the most likely face is larger than 0. Figure 2.3a illustrates this voting scheme.

2.1.1.5 Eye Localization Refinement

Our voting procedure results in a robust but rough localization of the eyes in the form of eye patches. In order to come up with an exact localization that can be used to apply the face image normalization procedure, we have to further refine this rough localization. Therefore we have trained an Active Appearance Model as described in Section 2.1. Since we are solely interested in the localization of the eyes, we restrict our model to incorporate only the eyes region (see Figure 2.4a,b for the setup of the eyes region).

The training data set consists of 427 manually annotated face images taken partly from the Caltech Faces database and partly from our own collection. Training images show variations with respect to open and closed eyes, eye gaze and only slight variations in head pose. During principal component analysis we keep 95 percent of the eigenvalue energy spectrum to represent our compact model on the highest resolution level. We use three levels of resolution and adapt the percentage of kept eigenvectors, i.e., on lower levels we restrict the shape variability to 90 and 80 percent respectively, while keeping the texture and appearance variability at 95 percent. This restriction enables a stronger focus on the global pose parameters for lower resolution levels.

AAM fitting is performed by initializing the mean shape of the AAM with the roughly estimated left and right eye locations from the facial component detection. This gives an initial solution for the pose parameters and the multi-resolution fitting algorithm is started with the lowest resolution. To switch between resolution levels it is necessary to upscale the shape and upsample the texture result and project it into the model of the higher resolution level. This gives the initialization for the higher level. After the fitting procedure has reached a local minimum we report the corresponding eyes shape. Additionally a confidence value derived from the final L_2 norm of the difference between synthetic model texture and warped image texture is calculated.

2.1.1.6 Face Normalization Transformation

The final step of our normalization work-flow is the calculation of the face normalization transformation and the resampling of the input image using this transformation according to the ICAO specification of eye locations for normalized token images [58]. Therefore we investigate the confidence value from the AAM model fitting and compare it to a fixed threshold that determines a measure of quality for the fitting. This fixed threshold was determined empirically on a set of typical test images. If the confidence value is large enough we take the center point between the AAM eye corners for left and right eye,

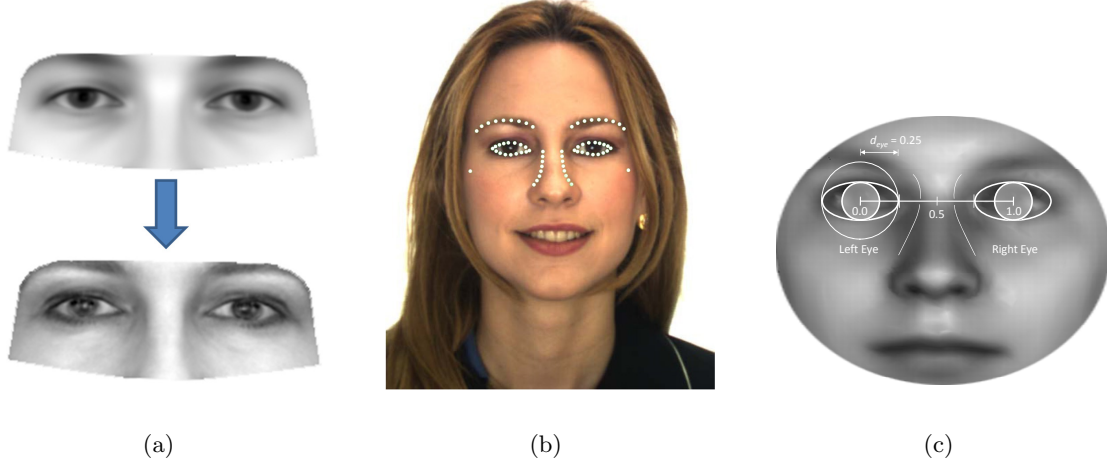


Figure 2.4: Detail illustration of the eye region. (a) Active Appearance Model: Learned mean shape/texture and the texture after successful fitting. (b) AAM shape model after successful fitting drawn on the input image. (c) Eye setup for the experimental evaluation measure *face detection rate*.

respectively, and compute the similarity transformation to map the eye coordinates to the standardized coordinates (see Figure 2.1). If the confidence in the AAM fitting is too low, we use the center points of the Boosting based eye localization patches as rougher estimates of the true eyes. Additionally if no face candidate with facial components was found, we report that no face was detected. Finally, the resampling is performed using bilinear interpolation. Note that for the evaluation it is only necessary to validate eye locations since the normalization is always the same similarity transformation.

2.1.2 Experimental Results

The accuracy of our facial image normalization procedure is tested on three different publicly available databases with known ground truth annotations of the eye centers. Since our intended application is face normalization we focus our evaluation on the eye centers. For evaluation we use the AR face database [90] consisting of 509 images, the BioID database [60] consisting of 1521 images, and the IMM database [103] consisting of 240 images. The databases show a variety of different people, poses, occlusions and facial expressions, and they have in common that there exists an annotation of the face intended for use in Active Shape Models where one can easily extract eye center locations from. Our face normalization pipeline leads to a left and right eye center location, we compare

this location with the ground truth annotation using the evaluation measure from [60]. We calculate the absolute pixel distance from the ground truth positions to receive two distance values. We choose the larger value and divide by the absolute pixel distance of the two manually set eye positions to become independent from face size according to [60]. We call this value *relative eye distance* d_{eye} . Further, we rate a face as found if the relative distance is equal or less than 0.25, which corresponds to an accuracy of about half the width of an eye in the image. The *face detection rate* is then calculated by dividing the number of correctly found faces by the total number of faces in the dataset (see Figure 2.4c). Images where no face can be found also contribute to this error measure as mis-detected faces.

We compare our method on all three databases against a standard Viola and Jones face detection approach using the implementation from [56]. For this standard implementation we use the default face detection parameters. We predict the eye location relative to the face rectangle using our large database of manually annotated eye rectangles by combining them with the Viola and Jones face detections. The *relative eye distance* error metric is presented in Figure 2.5 for the standard Viola and Jones algorithm and our method.

Additionally, we compare our results on the BioID database with the face detection results presented by the authors of the BioID study [60] (see Figure 2.5b), who call their approach *Hausdorff-MLP*. They show quantitative results solely on the BioID database report a *face detection rate* of 91.8% as opposed to 96.1% with our method. On the IMM and AR databases our *face detection rates* are 99.2% and 97.5%, respectively. Concerning the *face detection rate* we see that our method performs significantly better than the Hausdorff distance based method from [60]. This improvement is also reflected in the cumulative error distribution, where the ideal curve would be a step function to 100% at a relative eye distance of zero. The intersection of our proposed method with the relative eye distances of 0.05 and 0.1 are at approximately 62 and 89% compared to 39 and 79% for the *Hausdorff-MLP* method. From the *face detection rate* we can also see that the classic Viola-Jones method performs quite favorable in comparison to our method. However, after looking at the relative error distributions one can clearly see that the robust method shows a significant improvement in accuracy. This is very important since the definition of the face detection rate [60] allows a large amount of error. We conclude that our method outperforms the state of the art algorithms on these three publicly available data sets.

Finally, Figure 2.6 shows a number of qualitative results for the eye detection in the presence of pose deviations, facial expressions, and occlusions. The rightmost image of

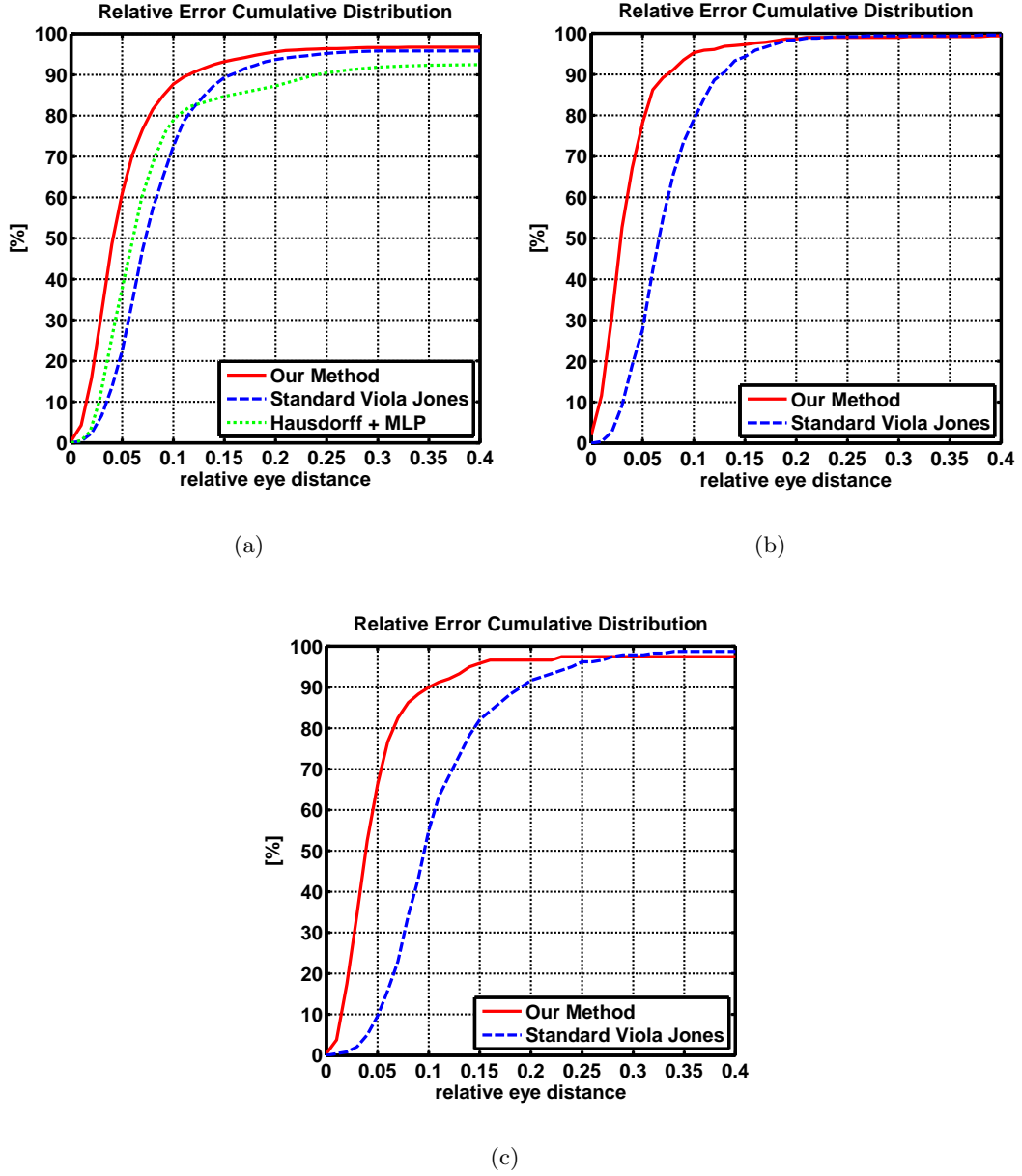


Figure 2.5: Quantitative results on relative error distributions of our method compared to different algorithms. (a) BioID database (1521 images), (b) AR database (509 images), (c) IMM database (240 images).

the bottom row shows a detection problem due to a pose deviation that is too large. Here also the symmetry detection has failed. The leftmost image of the bottom row shows an occluded eye, which is correctly detected by the probabilistic voting scheme. Images where no refined eye contour is drawn have resulted in a bad AAM fit, such that the center of the AdaBoost eye rectangles are used as eye centers. The other images show correct algorithm results. Finally, to assess algorithm run-time we did an evaluation on a small set of typical input images. The Viola-Jones implementation requires on average about 900 ms to detect a face from 800×600 images (note that we use the detection cascade on a large number of scales and with conservative parameter settings) on an Intel Core2 Duo notebook CPU at 2.53 GHz. Our algorithm performs of course slower due to the robust extensions. However, our run-time of on average about 3300 ms is still acceptable for our task of face image normalization.



Figure 2.6: Selected qualitative results on representative images from the databases.

2.1.3 Summary

In this section we presented a robust face normalization system suitable for transforming arbitrary images containing a face into a standardized coordinate system as, e.g., needed to fulfill the ICAO requirements on passport photographs. We show how our robust system is able to process input images with pose deviations, occlusions and over different facial expressions by making use of several redundant component detection methods. A robust approach is achieved by combining these detections. Our experimental results show superior results compared to state of the art algorithms. Further work is necessary to increase the range of pose deviations we are able to process. Here we need multi-view face and facial component detections which will be incorporated using multiple separately trained detection cascades over the possible range of poses.

2.2 Intensity-Based Congealing for Unsupervised Joint Image Alignment

For our AdaBoost based algorithms we need a lot of labeled training data. We manually labeled the facial images and align them in an unsupervised manner. Therefore we present an approach for the unsupervised alignment of an ensemble of images called congealing. Our algorithm is based on image registration using the mutual information measure as a cost function. The cost function is optimized by a standard gradient descent method in a multiresolution scheme. As opposed to other congealing methods [27, 28], which use the SSD measure, the mutual information measure is better suited as a similarity measure for registering images since no prior assumptions on the relation of intensities between images are required. We present alignment results on the MNIST handwritten digit database and on facial images obtained from the CVL database.

Congealing is the alignment of an ensemble of misaligned images. The only assumption in congealing is the type of geometric misalignment, e.g., translation, similarity, affine, and the assumption of a self-similar appearance class, e.g., faces, cars. There are several applications for congealing, e.g., the registration of a stack of images from different modalities in medical imaging [88] or the alignment of a training database for machine learning algorithms [53].

The seminal work of Learned-Miller [72] termed the notion “congealing”. They minimize parametric warp differences between a stack of images by applying a sum of entropies cost function. In recent work of Cox et al. [27] some problems of Learned-Miller are alleviated, namely the slow convergence, the need to select a stepsize and sensitivity on the warp parameterization. In their work, they applied a sum of squared differences (SSD) cost function to allow for an effective application of a Gauss-Newton gradient descent approach. They are able to simultaneously estimate warp parameter updates and they do not need a pre-defined step size as opposed to [72]. Their approach is similar to the well known Lucas & Kanade image alignment with the extension to an ensemble of images rather than a single image. Cox et al. further improved their results for a larger amount of images [28]. In their work, they claim that employing an inverse compositional formulation of least-squares congealing is superior to their additive formulation in [27] and thereby show an increase of alignment performance.

There are some other methods based on subspace techniques for automatically aligning an ensemble of images. Frey and Jojic [44] extended the Principal Component Analysis (PCA) to cope with non-aligned images. They obtained a set of aligned basis images

by applying the EM algorithm. However, one major drawback was the need to define a discrete set of allowable spatial warps affecting also computation time. De la Torre and Black [70] and Schweitzer [122] proposed extensions on Frey and Jojic’s approach. They learn a subspace, which is invariant to affine or higher order geometric transformations. The advantage is that the spatial warp variation is modeled continuously rather than discretely. The major drawback is the need for estimates of the basis images for the iterative algorithms which limits the applicability of their algorithms.

In this section we concentrate on the state-of-the-art congealing methods of [27, 28]. They make use of the SSD similarity measure. The SSD measure makes the implicit assumption that the images differ only by Gaussian noise after registration. Only in that case, the SSD measure is optimal [143]. For congealing this is never the case because we have a lot of intraclass variation, e.g., in the case of congealing facial images, different subjects, facial hair, gender or race. In other words, the SSD is not an appropriate cost function for generic image registration, hence we apply a more sophisticated cost function based on the mutual information measure [88, 133, 143]. This information-theoretic criterion is very general and powerful, because it does not depend on any assumption on the data (other than stationarity) and does not assume specific relations between intensities in a pair of images.

Based on the basic image registration method using mutual information we build our congealing approach, which is explained in detail in Section 2.2.1. Section 2.2.2 exhibits experiments congealing handwritten digit images and a stack of facial images. Our approach shows very good congealing results and is furthermore easy to implement. Finally, we discuss and summarize our work in Section 2.2.3.

2.2.1 Congealing

Congealing in our case is defined as an advanced case of image registration. In image registration we have one image called *moving image* $I_M(\mathbf{x})$ which is deformed to fit the other image, the *fixed image* $I_F(\mathbf{x})$. That is, we have to find a transformation $T_\theta(\mathbf{x})$ that aligns $I_M(\mathbf{x})$ with $I_F(\mathbf{x})$, where θ are the transformation parameters. The optimal transformation is found by minimizing a cost function \mathcal{C} with respect to θ

$$\hat{\theta} = \arg \min_{\theta} \mathcal{C}(\theta; I_F; I_M). \quad (2.1)$$

In the introductory section we noted that the SSD is not an appropriate cost function for generic image registration, hence we apply a more sophisticated cost function based

on the mutual information measure [88, 133, 143]:

$$\mathcal{C}(\theta; I_F; I_M) = - \sum_{m \in L_M} \sum_{f \in L_F} p(f, m; \theta) \log_2 \left(\frac{p(f, m; \theta)}{p_F(f; \theta) p_M(m; \theta)} \right), \quad (2.2)$$

where p is the discrete joint probability, p_F and p_M are the marginal probabilities, and L_F and L_M are sets of regularly spaced histogram bins containing intensity values of the fixed and moving image respectively. L_F and L_M together span a 2D joint discrete histogram $h(f, m; \theta)$ where the joint histogram values are estimated using Parzen windows w_F and w_M representing the fixed and moving image:

$$h(f, m; \theta) = \frac{1}{\sigma_F \sigma_M} \sum_{\mathbf{x}_i \in \Omega_F} w_F \left(\frac{f - I_F(\mathbf{x}_i)}{\sigma_F} \right) w_M \left(\frac{m - I_M(T_\theta(\mathbf{x}_i))}{\sigma_M} \right). \quad (2.3)$$

The scaling constants σ_F and σ_M must equal the intensity histogram bin widths defined by L_F and L_M . These follow directly from the grey-value ranges of I_F and I_M and the userspecified number of histogram bins $|L_F|$ and $|L_M|$.

The joint histogram $h(f, m; \theta)$ is proportional to the discrete joint probability $p(f, m; \theta)$ given by

$$p(f, m; \theta) = \frac{1}{|\Omega_F|} h(f, m; \theta) \quad (2.4)$$

where $|\Omega_F|$ is the number of pixels in the fixed image domain Ω_F . The marginal discrete probabilities p_F and p_M of the fixed and moving image are obtained by summing p over m and f , respectively

$$\begin{aligned} p(f; \theta) &= \sum_{m \in L_M} p(f, m; \theta) \\ p(m; \theta) &= \sum_{f \in L_F} p(f, m; \theta). \end{aligned} \quad (2.5)$$

The mutual information measure is very general; only a relation between the probability distributions of the intensities of the fixed and moving image is assumed. This cost function is minimized iteratively by a standard gradient descent method [67] in a multiresolution scheme. The advantage of a multiresolution scheme is to start the registration process at lower image complexity to reduce the sensitivity to get stuck in local minima of the cost function. Furthermore the overall runtime is decreased.

In (2.3) we observe a loop over pixel coordinates \mathbf{x}_i over the fixed image domain Ω_F . In general, it is not necessary to take all coordinates into account, but a smaller amount

of coordinates may already suffice [67, 133]. This subsampling strategy leads to a lower computational cost, especially for larger images. We use a random selection of a user-specified number of coordinates \mathbf{x}_i . Furthermore the sampling is performed by taking samples off the pixel grid to improve the smoothness of the cost function, as suggested by [81, 132].

The image registration functionality is provided by the *elastix* package [66], which also allows to choose some other cost functions, optimization techniques, subsampling strategies and interpolation methods.

For congealing we extend the concept of image registration. Every image of the ensemble of unaligned images is taken once as a moving image. This happens in an outer loop over all images. During one outer loop iteration all other images serve as fixed images. We register the moving image to every fixed image using the entropy based registration described above obtaining the transformation parameters θ . By averaging the transformation parameters obtained from all those registrations we get the final transformation parameters for the moving image. Note that this procedure is significantly simpler and easier to implement than the comparable approaches of [28, 72].

2.2.2 Experimental Results

First we evaluate our congealing algorithm on handwritten digits obtained from the MNIST database [73] in Section 2.2.2.1. The outcome of the experiments using handwritten digits motivated us to apply congealing to a more difficult object class. Therefore we show congealing with an ensemble of facial images in Section 2.2.2.2.

2.2.2.1 Congealing Handwritten Digits

We show the applicability of congealing using samples from the MNIST handwritten digit database [73]. A total of 50 randomly selected images per digit are used for our experiments. We allow an affine transform $T_\theta(\mathbf{x})$ for this image registration task having six parameters to optimize. The results are presented visually in terms of average images. Figure 2.7 (second row) shows the sharpness of the average images generated from congealing compared to the average images of the unaligned digits in Figure 2.7 (first row). It can clearly be seen that most of the spatial variation among the digits is removed. The average runtime* to congeal one sample of size 28x28 is 78s.

*The runtime is measured using an Intel Core 2 Duo processor running at 2.4GHz.

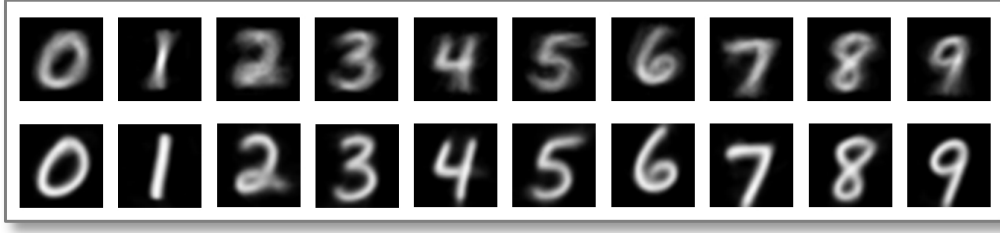


Figure 2.7: Average images before (first row) and after congealing (second row). The samples were obtained from the MNIST database [73].

2.2.2.2 Congealing Facial Images

Motivated from the results of our congealing experiments with handwritten digits in Section 2.2.2.1 we apply our algorithm also to facial images from the CVL database [109], because the images of this database show variation in gender, pose and facial expression and are not aligned. The CVL database consists of facial color images from 114 individuals of a resolution of 640x480 pixels. For our experiments we use the frontal pose image of every individual. We crop the faces to a size of 270x270 pixels and perturb the images randomly by a small amount of translation, scale and rotation to build a strongly unaligned set of facial images. In contrast to our first experiment in Section 2.2.2.1 we allow only a similarity transform $T_{\theta}(\mathbf{x})$ for image registration, because we do not want to shear our facial images. The unaligned facial images and the result of congealing are shown in Figure 2.9.

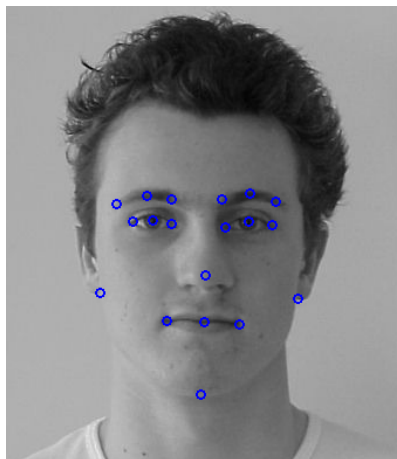


Figure 2.8: Annotation of a facial image with 19 landmark points at salient facial feature positions.

To be able to perform also a quantitative evaluation of congealing quality, we annotated all frontal images manually with our annotation tool (described in Appendix D) with 19 landmark points at salient facial feature positions, shown in Figure 2.8. We will provide our annotations for the public research community. We created a set of aligned landmarks used as groundtruth by applying Procrustes Analysis. The Point-to-Point distance of the unaligned landmarks (corresponding to the perturbed images) to the aligned landmarks is illustrated in Figure 2.10. After congealing we used the obtained landmarks and compared them also to the groundtruth exhibited in Figure 2.10. It can clearly be seen that the distribution of the Point-to-Point distances is shifted towards smaller displacements emphasizing the applicability of our algorithm. We also want to show the benefits of our mutual information cost function by replacing (2.2) by a SSD measure and compare the congealing results in Figure 2.10. The SSD is clearly outperformed by the mutual information measure, especially the SSD exhibits many outliers. These findings substantiate our claims from the introductory section.



Figure 2.9: The spatial variation gets removed from the perturbed facial samples by our algorithm. The samples are taken from the CVL database [109].

2.2.3 Summary

We presented an algorithm for the unsupervised alignment of a stack of images. The commonly used SSD measure for congealing is not appropriate for generic image registration. Hence, we used the more sophisticated mutual information similarity measure as a cost

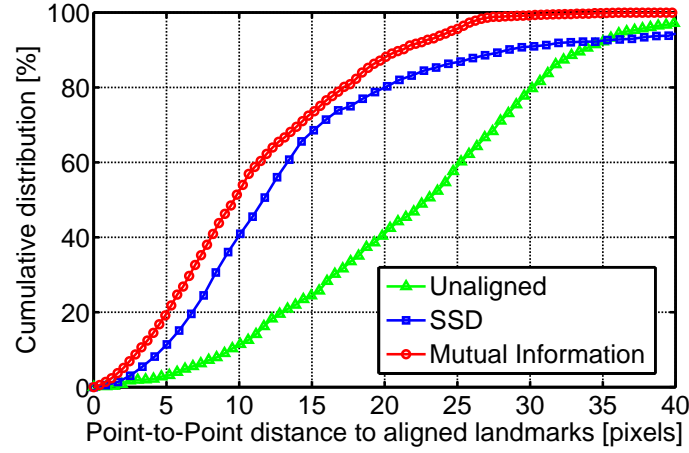


Figure 2.10: Point-to-Point distance of the unaligned landmarks and the congealed landmarks to the aligned landmarks.

function. This cost function is optimized by a standard gradient descent method in a multiresolution scheme. The congealing results on the MNIST handwritten digit database and the results for congealing facial images obtained from the CVL database clearly show the applicability of our algorithm. We also provide our annotations on facial images of the CVL face database for the public research community.

Eyes and Mouth Analysis

After the detection of eyes and mouth components in a facial image, an analysis procedure is applied to assess the ICAO criteria *eyes-open* and *mouth-closed*. Reviewing the literature revealed a multitude of techniques being applied to this problem, among them machine-learning approaches like support vector machines [139], or Boosting [120], model-based approaches like EigenFaces [137] or Active Appearance Models [22], or simpler geometric and template based methods for detecting eye- or mouth-related features like lips, teeth, iris, or eyelids. Despite their usefulness in many situations, all of these approaches have their specific drawbacks, e.g., performance of model-based approaches decreases significantly in the presence of outliers, while their accuracy is superior to other methods if the model fitting is successful. From the observation of differing performance of different - to a certain extent complementary - algorithms, we adapted the interpretation of each algorithm as a single expert giving a vote for a certain classification decision. By combining the votes of all classifiers in a classifier fusion scheme [65, 69], we state the hypothesis that the performance of the combined scheme is superior compared to the performance of the single classifiers in the ensemble. This makes the decision for the specific events *eyes-open* and *mouth-closed* more robust in the presence of difficult situations like noisy input data, lighting conditions or partial occlusions, e.g., wearing glasses.

In the following we will validate our hypothesis by presenting our face analysis system consisting of the single classifiers (Section 3.1) and the classifier fusion strategies in more detail in Section 3.2. Section 3.3 shows the results of our experiments on two databases. Finally we discuss and summarize our findings in Section 3.4.

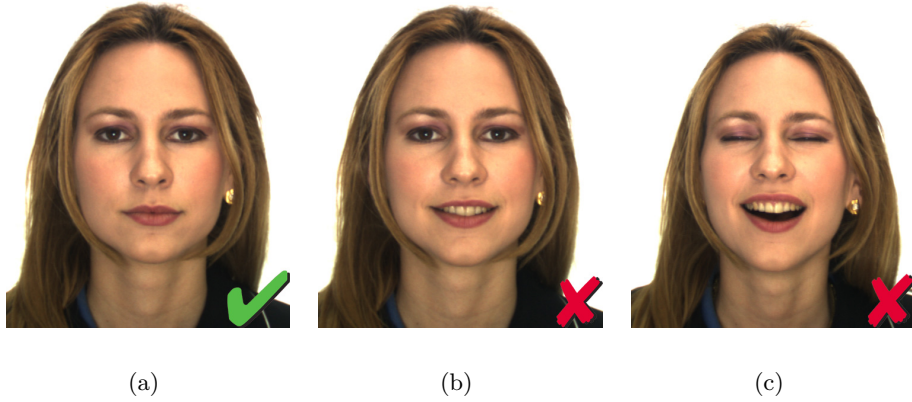


Figure 3.1: Some sample images from the evaluation database: (a) eyes open, mouth closed, (b) eyes open, mouth open, (c) eyes closed, mouth open. Taken from the AR face database [90].

3.1 Face Analysis System Description

Our face analysis system operates on *tokenized* images. It performs several classification decisions of which we present the *eyes-open* and *mouth-closed* events in this section. These criteria rely on our facial component detection stage, where a robust scheme performs face, eye and mouth localization from face component hypotheses in a probabilistic voting framework, see Section 2.

The ICAO specification [57] defines the following rules for accepting photos as suitable according to *eyes-open* and *mouth-closed* criteria. The face expression has to be neutral (non-smiling) with both eyes open normally and the mouth closed (see Figure 3.1). A smile is unacceptable regardless of the inside of the mouth and/or teeth being exposed or not. Starting from this specification and taking the large variety of possible problems in real-world images (due to noisy data, inappropriate lighting situations, occlusions due to hair or glasses, or the large variety in appearance of different people) it is intuitive that a single classification method will not be able to solve this task in a robust manner. Therefore several different classifiers are combined in a fusion step (see Figure 3.2). An important assumption for efficiently combining classifiers is that they show complementary behavior and their estimates are as independent as possible. In practice it is very hard to come up with a set of totally independent methods, so one has to rely on experimental evaluation to show their applicability to a given task.

For the training based approaches we have used a large manually annotated training

set of around 4600 face images which were taken from the Caltech Face database [19], the FERET database [111] and from a third database constructed from our own images.

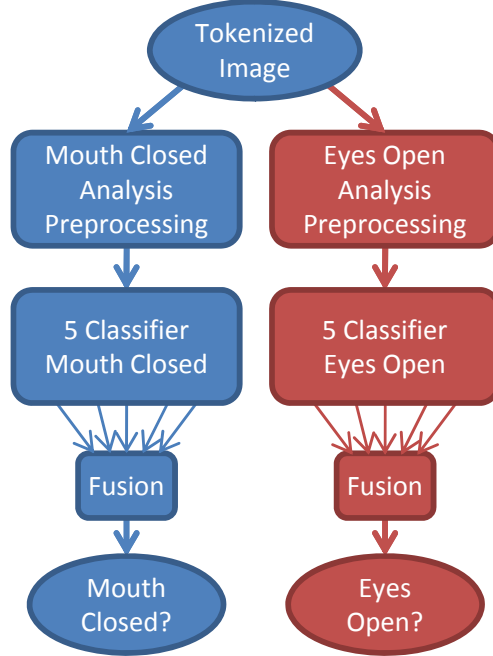


Figure 3.2: Face analyzer workflow. From tokenized images we perform some pre-processing, apply the single classifiers and fuse their results to form a final decision.

3.1.1 Eyes Open Analysis

For the analysis of the event *eyes-open* we use an ensemble of five classifiers. Two classifiers are based on AdaBoost, one uses the Active Appearance Model, one is derived from the *EigenFace* method and the last method is based on a geometric iris localization strategy. The *eyes-open* decision is performed independently for the left and right eye and leads to a confidence value $d_i(\mathbf{x}) \in [0, 1]$ representing the range between closed and open eyes. The minimum of these two separate decisions forms the final result, since one closed eye already corresponds to an *eyes-closed* decision.

3.1.1.1 Active Appearance Model

We trained an Active Appearance Model (AAM) [22], (Appendix B), (CL1), for face image regions around the eyes, which we have also used in the tokenization step, see Figure 2.4. Our training set consists of 427 manually annotated face images taken from

the Caltech Face database and our own collection. Training images show variations in the opening of the eyes, slight pose variations, and eyes, which are looking straight and away. For model building we keep 97.5 percent of the eigenvalue energy spectrum to represent our compact model. To apply the AAM to a given image for *eyes-open* classification, we initialize the mean shape of the AAM by the roughly estimated left and right eye locations from facial component detection. To derive a measure of the likelihood of the *eyes-open* event we analyze the vertical eyes' opening of the converged AAM shape model in the left and right eye area respectively. We compare the opening to a pre-defined threshold $T_{E,aam}$ and additionally weight the distance to the threshold with the AAM residual error that represents an estimate of success or failure of model fitting.

3.1.1.2 Iris Detection Approach

Our geometric iris detection approach (**CL2**) is based on a fast radial symmetry detector presented in [86]. For each eye we restrict ourselves to an image patch around the eye. After performing edge-preserving anisotropic smoothing [145], we calculate a symmetry transform image by estimating gradient orientation and magnitude projection images over several scales according to [86]. Local minima of the symmetry transform image correspond to centers of radial-symmetric structures. The strongest response of this transform corresponds to iris centers. Afterwards we perform a more accurate iris radius estimation by using a one-dimensional Hough voting on the binary response image from a Canny edge detector [20]. We favor iris radii that are conform to a rough scale estimation of the iris that we are able to derive from our tokenized input images. The voting histogram entry with the maximal response gives the desired iris radius. From the strength of the symmetry image minimum and the voting histogram we derive a confidence measure for the *eyes-open* event.

3.1.1.3 AdaBoost Classifier

Eyes-open analysis using AdaBoost [120], [142], (Appendix C), utilizes two different classifiers, both trained with the OpenCV [56] library. These classifiers focus on Haar wavelet filter features. The first one (**CL3**) was trained on image patches of closed eyes and the second (**CL4**) on open eyes. For the closed eye classifier 464 positive image patches were used, while the open eye classifier was trained with 2732 image patches. The discrepancy in the number of positives is due to the unequal representation of both classes in our training set. In both cases the set of negative images was taken from generic

background images.

Our classification strategy for each trained Boosting classifier takes the approximate location of the eye from the facial component detection stage and applies the classifier to a slightly enlarged region around this region. That is the reason, why we trained one open-eye and one closed-eye detector utilizing the sliding window approach in the enlarged region, hence exploiting the detector as a classifier. If we detect an open eye rectangle we report a confidence measure according to [149]. If we detect a closed eye rectangle we report the inverse of this confidence measure.

3.1.1.4 EigenEye Model

The *EigenFace* model [137] is a generative method that explicitly models face images in the face image space. It is based on registered image patches of equal dimensions and is calculated by applying a PCA to the face image vectors in the face image space. From this compact PCA representation, the eigenvectors according to the smallest eigenvalues are discarded. Given such a model consisting of a mean image and the remaining eigenvectors, it is possible to represent unknown image patches in the face image space. Thus, the reconstruction error of an image patch gives information about the similarity to facial images. By applying a threshold on the reconstruction error, classification can be performed.

The same principle can be used to model open eyes (*EigenEye*, **CL5**). This generative *EigenEye* model is based on a training set of approximately 2700 open eye images for left and right eye, respectively. These images are coarsely registered before model calculation. From the compact PCA representation we keep 120 eigenvectors, resembling 97.5 percent of the energy in the eigenvalue spectrum. For deciding on the similarity to the space of open eyes we define a threshold $T_{E,pca}$ on the reconstruction error. The distance to this threshold gives a continuous confidence measure.

3.1.2 Mouth Closed Analysis

For the *mouth-closed* analysis we have also used an ensemble of five classifiers. Three classifiers are based on AdaBoost, one approach makes use of the *EigenFace* framework and the last classifier utilizes a blob detection algorithm that locates dark blobs due to mouth cavity shadows. Decision scores $d_i(\mathbf{x}) \in [0, 1]$ range between 0 for open and 1 for closed mouths.

3.1.2.1 Geometric Dark Blob Analysis

The dark blob analysis (**CL1**) is a geometric method that makes use of the fact that open mouths very often exhibit dark blobs due to shadows in the open mouth cavity compared to the rest of a mouth image patch. Therefore, we investigate a slightly enlarged version of the mouth detection area, transform it into HSV color space and proceed by working solely on the Value coordinate. After binarizing the mouth patch using thresholding [105], we perform a blob detection process that extracts dark blobs corresponding to shadow regions. A filtering stage on the extracted blobs regarding their size, center locations and compactness removes unlikely shadow regions that may occur, e.g., due to beards. If a dark blob region survives this filtering stage we decide for the *mouth-open* event, otherwise for *mouth-closed*. A confidence measure is derived from the size of the detected blob region.

3.1.2.2 Boosting Classifier

Mouth-closed analysis using AdaBoost (Appendix C) leads to three different classifiers. The first one (**CL2**) is trained with the OpenCV library using 3785 closed mouth patches as positives and a large pool of non-mouth patches as negatives. This classifier focuses on Haar wavelet filter features. The classification strategy takes the approximate location of the mouth from the facial component detection stage into account and applies the classifier to a slightly enlarged region around the mouth.

The second AdaBoost classifier (**CL4**) uses integral image approximations of edge orientation histograms. The weight update strategy follows the *RealBoost* scheme. We expect complementary behavior of our *RealBoost* approach to the OpenCV implementation due to the different features under consideration, i.e., the OpenCV library focuses on wavelet filter approximations, while our *RealBoost* learns features from the edge information of an image. *RealBoost* is applied similar to the first classifier but uses 2475 open mouth patches as positive images in the training stage.

The third AdaBoost classifier (**CL3**) is also trained with the *RealBoost* scheme on 1200 closed mouth patches. This classifier uses the same amount of open mouth patches as negatives and can only be applied directly (without a sliding window approach) to the detected mouth patches from the facial component detection. All of the AdaBoost classifiers report a confidence measure which is calculated according to [149] in case of closed mouths.

3.1.2.3 EigenMouth Model

Similar to the *EigenEye* model, the generative *EigenMouth* model (**CL5**) is based on a training set of 3785 closed mouth images in neutral expression. These images are coarsely registered and the model is calculated. From the compact PCA representation we keep 140 eigenvectors, resembling 97.5 percent of the energy in the eigenvalue spectrum. For deciding on the similarity to the space of closed mouths we define a threshold $T_{M,pca}$ on the reconstruction error. The distance to this threshold gives a continuous confidence measure.

3.2 Classifier Fusion

We hypothesize that fusing multiple classifiers generates more accurate classification results compared to single classifier decisions. Hence, our goal is to evaluate different fusion strategies to combine the classifiers discussed in the previous sections. Let D denote a single classifier and $\mathbf{x} \in \mathbb{R}^n$ a feature vector representing a pattern to be classified. The classifier represents a mapping

$$D : \mathbf{x} \in \mathbb{R}^n \rightarrow \omega_j \in \Omega,$$

where ω_j is one of the c possible classes of $\Omega = \{\omega_1, \dots, \omega_c\}$. Denote $\{D_1, \dots, D_L\}$ as the set of L classifiers. The output of the i th classifier is $D_i(\mathbf{x}_i) = [d_{i,1}(\mathbf{x}_i), \dots, d_{i,c}(\mathbf{x}_i)]^T$, where \mathbf{x}_i is the specific feature vector representation of the input pattern needed by classifier D_i and $d_{i,j}(\mathbf{x}_i)$ is the confidence, i.e., the degree of support, classifier D_i assigns to the assumption of \mathbf{x}_i originating from class j . The fused output \hat{D} of the L single classifiers is

$$\hat{D}(\mathbf{x}) = \mathcal{F}(D_1(\mathbf{x}), \dots, D_L(\mathbf{x})), \quad (3.1)$$

where \mathcal{F} is called the *fusion strategy*. Resulting from \hat{D} , the final confidence values assigned to each class are \hat{d}_j . The following fusion strategies are investigated:

Majority vote (MAJ) The outputs of the single classifiers $d_{i,j}$ are assigned to a specific class explicitly. The class label that occurs most often is taken as the final decision.

Minimum (MIN) $\hat{d}_j(\mathbf{x}) = \min_i \{d_{i,j}(\mathbf{x})\}$

Maximum (MAX) $\hat{d}_j(\mathbf{x}) = \max_i \{d_{i,j}(\mathbf{x})\}$

Average (AVR) $\hat{d}_j(\mathbf{x}) = \frac{1}{L} \sum_{i=1}^L d_{i,j}(\mathbf{x})$

Binarized Average (BAVR) This scheme is equivalent to the average fusion strategy, except for the outputs of the single classifiers $d_{i,j}$ being assigned to a specific class explicitly before averaging.

Product (PRO) $\hat{d}_j(\mathbf{x}) = \prod_{i=1}^L d_{i,j}(\mathbf{x})$

Prior Confidence (PRIOR) A priori confidences of the single classifiers are obtained from tests on a validation dataset according to their performance exhibited (Table 3.1). The prior confidences are accumulated according to the decision of the corresponding single classifier.

Bayes Combination (BAYES) This scheme assumes that the classifiers are mutually independent and that the posterior confidences of the single classifiers are equal to posterior probabilities. For our single classifiers we expect independence, because different underlying concepts and methodologies are used, e.g., different features for classification.

The fusion strategies presented above can only be justified under strict probabilistic conditions. Nevertheless, some of them exhibit excellent performance as can be seen in the experimental results, a fact which was already stated in [65] for some of the classifier fusion strategies. The strict probabilistic conditions motivate to learn the fusion of the single classifiers. Thus, we use a support vector machine (SVM) [139], [82] to learn the optimal fusion in a linear combination.

SVM learned posterior confidence fusion (SVM POST) Obtained posterior confidences on a labeled validation dataset (independent of the training- and test set) are used for the learning of a linear combination of the single classifiers by the use of a SVM.

SVM learned binary posterior confidence fusion (SVM BPOST) Obtained posterior confidences on a validation dataset are assigned to a specific class explicitly. These binarized confidences are then used for learning the optimal combination by the use of a SVM.

SVM learned Bayes combination (SVM BAYES) In this scheme the optimal Bayes combination of the single classifiers confidences is learned by the use of a SVM on a validation dataset.

3.3 Experimental Results

We used two different face databases for the evaluation of our single classifiers and fusion strategies. The first one is the publicly available **AR face database** [90]. It consists of 126 unique people, frontal view face images (70 male, 56 female) of different facial expressions, illumination conditions and occlusions resulting in a total amount of over 4000 color images of a resolution of 768 by 576. We used all the images except those with dark sunglasses or occluded mouth due to a scarf yielding about 1700 images for evaluation. Annotation data is available on request. Some samples of this challenging database are given in Figure 3.3. The **second database** we used is a private data set containing 325 frontal face color images of 480 by 640 pixels with 30 people showing different facial expressions.



Figure 3.3: Sample images from the AR face database [90] showing the difficulties of the images under consideration.

The evaluation procedure is exemplified on our own database for the AAM *eyes-open* classifier in Figure 3.4. All evaluation results on the AR face database are summarized in Table 3.2 for *eyes-open* and in Table 3.4 for *mouth-closed* analysis. The evaluation results on our private database are presented in Table 3.3 for *eyes-open* and Table 3.5 for *mouth-closed* analysis. The comparison of the ROC curves of the best single classifier to the best fusion strategy is shown in Figure 3.5 for the AR face database and in Figure 3.6 for our own database.

The best overall fusion performance is exhibited by the learning based SVM approaches, and here especially the SVM using the posterior confidences. However, also

Table 3.1: Prior confidences of the single classifiers. (left) eyes-open classifiers, (right) mouth-closed classifiers

Classifier	Prior	Classifier	Prior
CL1	0.25	CL1	0.23
CL2	0.20	CL2	0.21
CL3	0.25	CL3	0.25
CL4	0.19	CL4	0.21
CL5	0.11	CL5	0.10

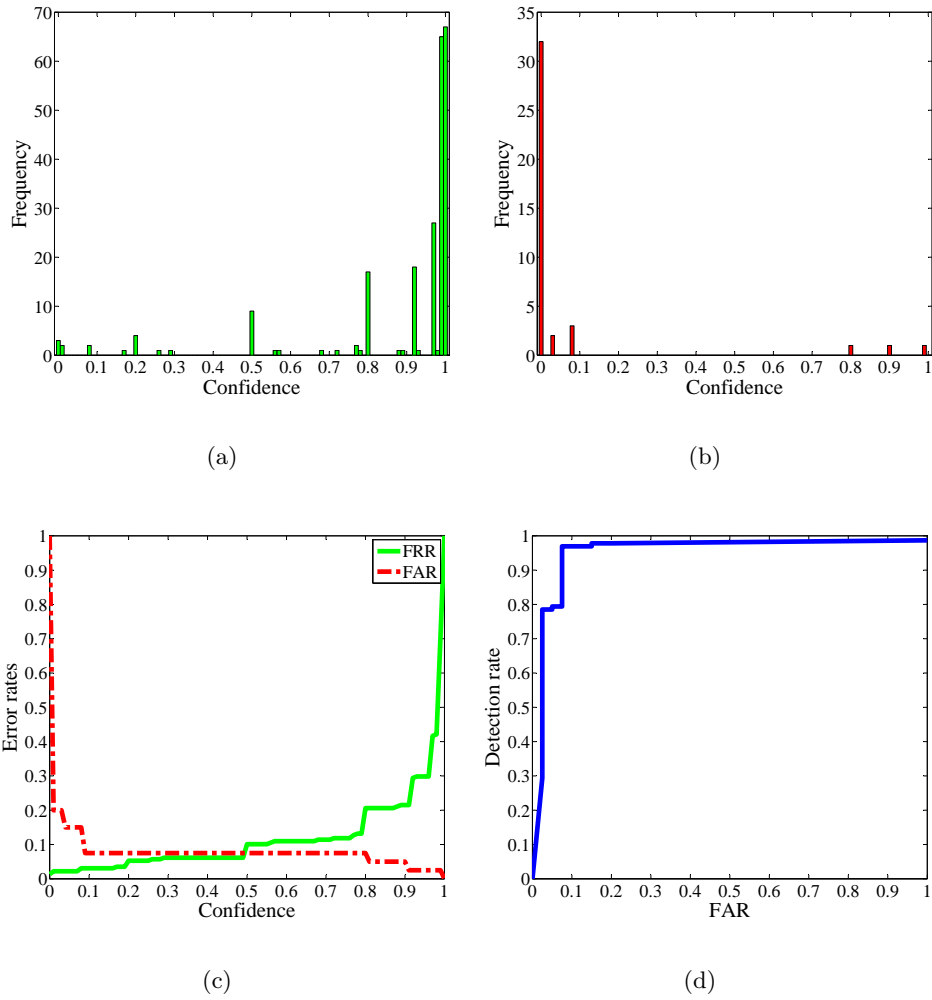


Figure 3.4: Evaluation procedure for the eyes-open AAM on our own database. (a) Distribution of the confidence values for open-eyes in the database. (b) Distribution of the confidence values for closed-eyes in the database. (c) Characteristic of the false rejection rate (FRR) and false acceptance rate (FAR). (d) ROC curve

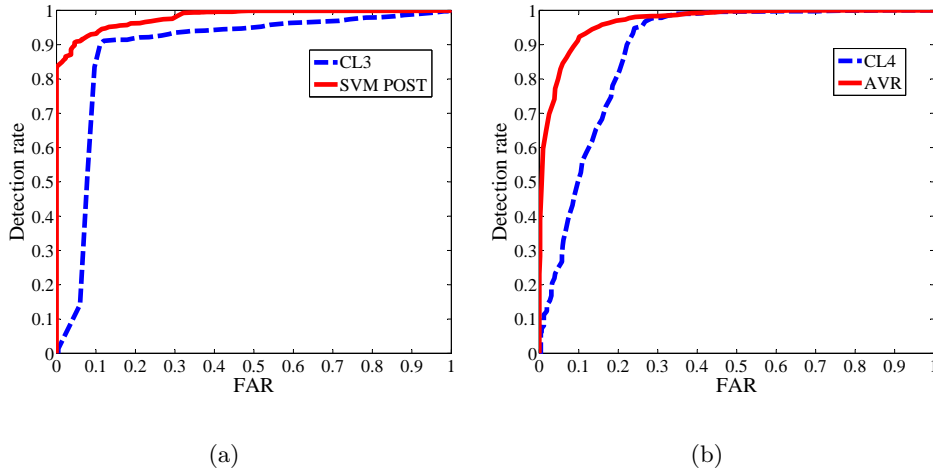


Figure 3.5: Comparison of the ROC curves of the best single classifier to the best fusion strategy for the (a) eyes-open and (b) mouth-closed analysis on the AR face database.

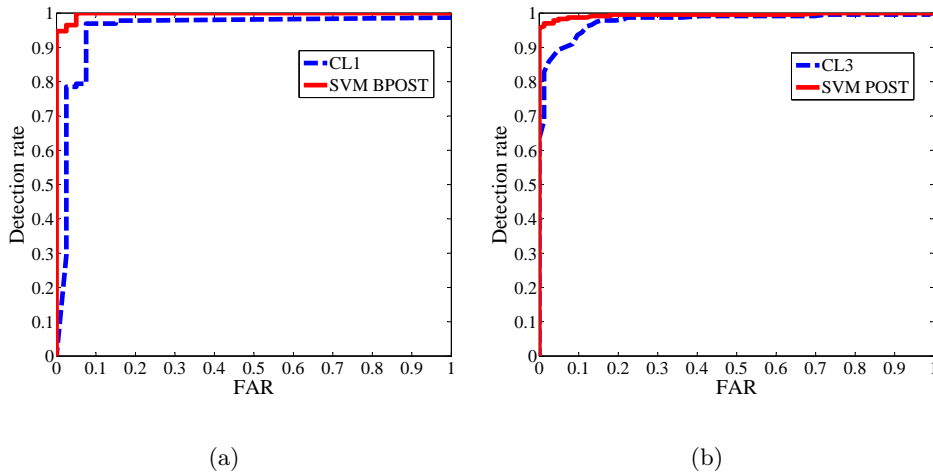


Figure 3.6: Comparison of the ROC curves of the best single classifier to the best fusion strategy for the (a) eyes-open and (b) mouth-closed analysis on our own face database.

simple fusion strategies, e.g., average, show very good results. The single classifier performance on our private database is slightly better compared to the AR database due to a larger complexity of the images contained in the latter data set. The AR face database contains a large number of images from people wearing different glasses, often show severe specular reflections, simulate bad illumination conditions and extremely wide open mouths, where we have problems of robustly locating the mouth region. However, the hypothesis of fusing multiple classifiers generating more accurate and robust classification

results compared to a single classifier, is thus approved, illustrated in our figures and tables.

Table 3.2: Evaluation results for the eyes open analysis on the AR face database

	EER	Detection Rate @ FAR			
		0.05	0.1	0.3	0.5
CL1	18.35	48.90	68.68	86.66	94.14
CL2	29.37	70.63	70.63	70.63	70.63
CL3	10.85	11.96	85.18	93.49	95.04
CL4	24.89	74.45	74.61	75.37	76.65
CL5	25.75	47.87	63.86	76.04	81.12
MAJ	18.08	81.92	81.92	81.92	81.92
MIN	26.91	71.38	71.77	73.34	75.44
MAX	15.91	29.00	62.17	99.13	99.73
AVR	9.21	87.96	91.17	96.23	99.66
BAVR	7.82	81.92	92.18	92.18	99.73
PRO	28.03	70.59	70.89	72.09	73.28
PRIOR	7.82	90.55	92.18	92.18	99.73
BAYES	8.81	89.22	91.65	96.99	99.82
SVM POST	7.65	90.86	93.19	97.97	99.86
SVM BPOST	8.26	85.72	92.11	99.25	99.73
SVM BAYES	11.31	85.18	87.91	96.03	99.86

Table 3.3: Evaluation results for the eyes open analysis on our own face database

	EER	Detection Rate @ FAR			
		0.05	0.1	0.3	0.5
CL1	7.50	79.39	96.93	97.92	98.14
CL2	10.96	89.04	89.04	89.04	89.04
CL3	13.51	85.09	85.53	92.65	94.88
CL4	12.72	87.28	87.28	87.28	87.28
CL5	15.30	67.11	83.77	88.16	90.94
MAJ	5.70	94.30	94.30	94.30	94.30
MIN	13.16	86.84	86.84	96.38	97.04
MAX	6.50	87.72	99.56	100.00	100.00
AVR	3.95	96.05	96.49	100.00	100.00
BAVR	5.00	96.05	96.05	100.00	100.00
PRO	12.97	86.64	86.89	87.85	88.82
PRIOR	3.95	96.05	99.56	100.00	100.00
BAYES	3.73	96.49	96.49	100.00	100.00
SVM POST	3.51	96.49	98.25	100.00	100.00
SVM BPOST	3.51	100.00	100.00	100.00	100.00
SVM BAYES	5.00	95.18	98.25	100.00	100.00

Table 3.4: Evaluation results for the mouth closed analysis on the AR face database

	EER	Detection Rate @ FAR			
		0.05	0.1	0.3	0.5
CL1	36.90	0.00	0.00	0.00	92.31
CL2	28.33	64.83	66.03	72.18	80.17
CL3	25.13	31.49	44.88	79.67	95.31
CL4	19.53	25.34	50.80	97.83	99.67
CL5	23.90	47.66	60.63	81.36	90.31
MAJ	17.12	24.20	48.39	97.16	97.16
MIN	16.91	60.36	70.61	89.66	89.76
MAX	39.77	42.49	43.16	45.82	98.26
AVR	9.40	81.12	91.90	98.33	99.75
BAVR	15.71	42.94	84.29	97.16	99.50
PRO	11.78	75.79	85.96	93.60	94.64
PRIOR	14.34	50.63	84.29	97.41	99.67
BAYES	10.79	79.18	88.33	98.07	99.75
SVM POST	11.21	73.74	87.45	97.17	99.75
SVM BPOST	14.14	51.53	63.99	98.08	99.67
SVM BAYES	13.35	67.94	83.06	93.57	99.67

Table 3.5: Evaluation results for the mouth closed analysis our own face database

	EER	Detection Rate @ FAR			
		0.05	0.1	0.3	0.5
CL1	12.02	87.98	87.98	87.98	87.98
CL2	11.86	53.48	85.49	94.97	95.78
CL3	8.62	89.36	93.82	98.71	99.14
CL4	12.39	55.56	81.93	98.02	99.57
CL5	25.50	25.17	51.52	77.21	84.55
MAJ	3.61	97.00	97.00	97.00	97.00
MIN	16.03	83.76	83.86	84.24	84.62
MAX	13.04	40.26	47.47	100.00	100.00
AVR	3.61	97.85	97.85	99.14	100.00
BAVR	3.61	97.00	97.00	98.71	98.71
PRO	8.81	91.00	91.24	92.24	93.23
PRIOR	3.00	97.00	98.71	99.57	99.57
BAYES	2.26	97.85	98.82	99.57	99.57
SVM POST	3.00	98.28	98.71	99.57	99.57
SVM BPOST	3.00	97.00	98.71	99.57	99.57
SVM BAYES	3.11	97.67	98.78	99.57	99.57

3.4 Summary

In this section we presented our face analysis system for checking ICAO specification compliance regarding the *eyes-open* and *mouth-closed* criteria. Both criteria are challenging due to different possible input images showing noise, bad lighting, occlusions (glasses, beard) and the large variety of human faces in general. Our face analysis system builds upon several pre-processing steps before *eyes-open* and *mouth-closed* events are evaluated by different, complementary classification methods. We adopt a classifier fusion framework testing a number of different fusion strategies to combine the votes of different classifiers. In our experimental results we show that the classification performance on our criteria improves significantly due to classifier fusion, thus validating our hypothesis. The fusion strategies based on the learned linear combination utilizing the SVM approach show superior results, however, simpler fusion strategies exhibit competitive performance.

Analysis of the Deviation from Frontal Pose

Automatically determining the head orientation from images has been a challenging task for the computer science community for decades [100]. The head can be assumed to be modeled as a rigid object, and therefore the pose of the head can be characterized by *pitch*, *roll* and *yaw* angles as illustrated in Figure 4.1. Head pose is also essential for understanding the eye's gaze, i.e., to determine the viewing direction of a person. In [71] it is shown, that the gaze direction is a combination of head pose and eye's gaze.

Many exciting and important application areas for head pose estimation emerged in the last decade. One major interest is human-computer interaction to determine the gaze as well as interpreting head gesturing, i.e., the meaning of head movements like nodding. This enables very direct means to interact with virtual worlds, especially in the computer gaming industry. A second area of application is automotive safety. To avoid vehicle collisions, the driver's head is monitored to recognize driver distraction and inattention. A 3D head tracking approach for a driver assistance system is presented in [99]. Biometrics also utilizes the important task of head pose estimation. One direction of impact is the rectification of non-frontal facial images to the frontal pose to improve the accuracy of face recognition [14]. Recently, extraordinary attempts to person surveillance for far-field distance views were announced. Even though there is a broad area of applications and a high demand for accurate systems, research on identity-invariant head pose estimation shows fewer evaluated systems and generic solutions compared to face detection and face recognition.

Murphy-Chutorian and Trivedi [100] give a recent and extensive survey on head pose

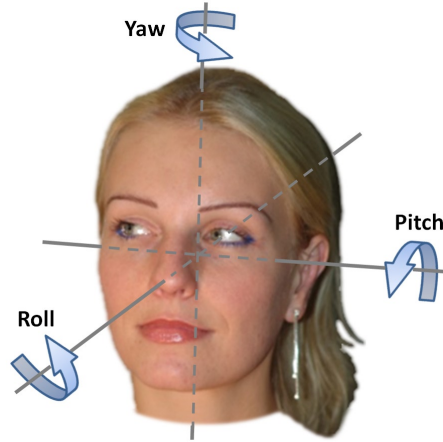


Figure 4.1: The three degrees of freedom of a human head described by *pitch*, *roll* and *yaw* angles.

estimation in computer vision. They arrange the methods published over the last 14 years in several categories: appearance template based methods, detector arrays, manifold embedding, flexible models, geometric methods, tracking methods and hybrid methods:

Appearance Template Methods and Detector Array Methods try to estimate the head pose by either directly comparing head images with a set of template images or by using a trained detector in order to find the most similar pose for a new head image.

Regression Methods learn a continuous estimate of the head pose by a possibly non-linear mapping of the high dimensional image features to pose angles.

Manifold Embedding Methods assume that even though an image consists of hundreds of dimensions spanned by the pixels of the image, only a few dimensions define the pose [92]. Thus these methods map the head image to a low-dimensional manifold that is defined by the continuous pose angles. Both methods rely on a robust and accurate face localization.

Geometric Methods rely more on human perception and include measurements of distances between facial features and deviation from bilateral symmetry [146]. These methods require that facial features such as eye corners can be detected robustly. Due to occlusion, facial expressions and pose-dependent viewing angles this is a non-trivial task [140].

Flexible Models solve this problem by fitting non-rigid face models to the 2D image of a face in order to determine the head pose. For example, graphs of facial features can be automatically deformed until they fit to a face in an arbitrary pose [147].

In contrast to many other approaches we are specifically interested in the more difficult task of head pose estimation from still images, instead of tracking faces in image sequences. Head pose estimation from still images is a challenging task due to the high variability of facial appearance, e.g., gender, ethnicity or age. A Head Pose Estimation System (HPES) should demonstrate robustness to such factors and also to occlusion, noise, lighting and perspective distortion. In the context of ICAO type passport photos, persons are required to show a frontal head pose which means that the head rotation must not deviate more than ± 5 degrees in any direction from frontal.

In this section we propose three head pose estimation approaches starting from a coarse estimation (only frontal/non-frontal decision) to a very fine estimation. The coarsest approach (Section 4.1) is based on AdaBoost classifiers and only performs a frontal vs. non-frontal decision. The next finer estimation approach (Section 4.2) builds upon a Histogram of Oriented Gradients (HOG) descriptor used as input for a non-linear regression. Furthermore a Biased Manifold Embedding (BME) approach is extended to cope with multiple pose-angles. In addition, we present an approach for the creation of an artificial training database. The finest head pose estimation approach is the 3D Morphable Appearance Model which is proposed in Section 4.3. The finer head pose estimation approaches are more accurate but are only applicable in a smaller range of head poses.

4.1 Boosting Approach

The analysis of the deviation from frontal pose is carried out using a combination of six different AdaBoost classifiers (Appendix C) using Haar wavelet like features. The classifiers make a decision on frontal-, left-, right-, up- or down pose deviation (Figure 4.2). There is one classifier which is trained on upwards-looking versus frontal pose face images, another one trained on downwards-looking versus frontal face images. For the deviations from left- and right pose we combine four different classifiers. The first classifier is trained using centered cropped face images, the second classifier takes face images, which are cropped taking a shifted window around the face center, see Figure 4.3. In order to incorporate symmetry information in the third classifier, the original image and its mirrored image are overlaid and the average pixel value is taken to form a new image, see Figure 4.4.

The advantage is that deviations from frontal pose become “amplified”. So the boosting approach is able to discriminate better between pose- and frontal images. This classifier can only determine, either if the face is frontal or not. Hence, a fourth classifier is needed to decide between left- or right pose, see Figure 4.2.

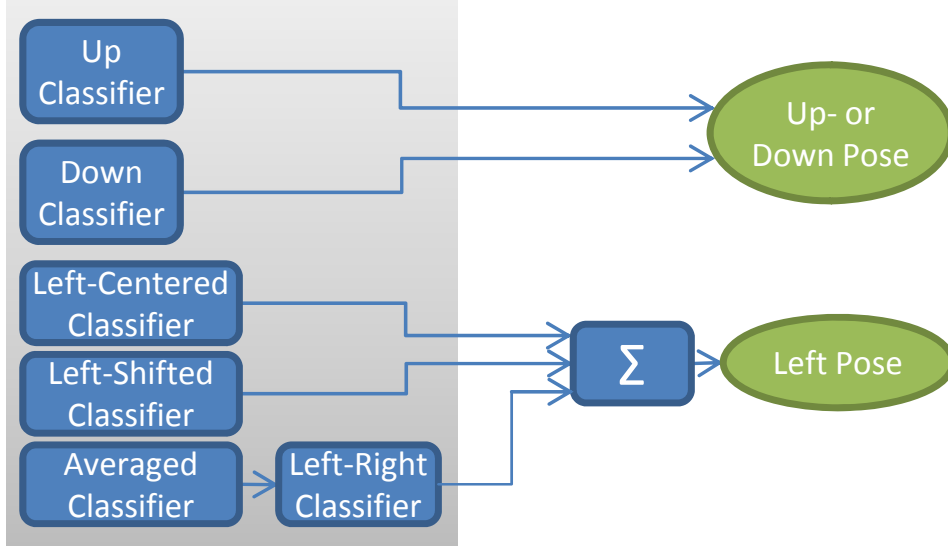


Figure 4.2: Pose classification with six AdaBoost learned pose classifiers.

4.1.1 Experimental Results

The performance of the AdaBoost based head pose estimation algorithm is evaluated on our own database. This database consists of 1355 JPEG-compressed face images of 35 individuals. Various scenarios, e.g., mouth open/closed, eyes open/closed, frontal/left/right pose, different eye gaze directions, partial face occlusion, are covered equally throughout the database. Note that this database is different from the data that we use for determining the prior confidences of our multi-classifier fusion step and different from the dataset we use for the training of the AdaBoost classifiers. Table 4.1 gives a summary of the prior confidences for combining the classifiers. The applied fusion strategy is a linear combination of prior confidences and AdaBoost classifier output confidences.

Our results are compared to the state-of-the-art technology of two commercial vendors, which we are unfortunately not allowed to mention, therefore we refer to them as vendor 'X' and vendor 'Y'. Figure 4.5 shows the comparisons of the relative detection rate for the analysis of the deviation from frontal pose. We show our results using the best performing system as the reference to which the other systems are compared. This rela-

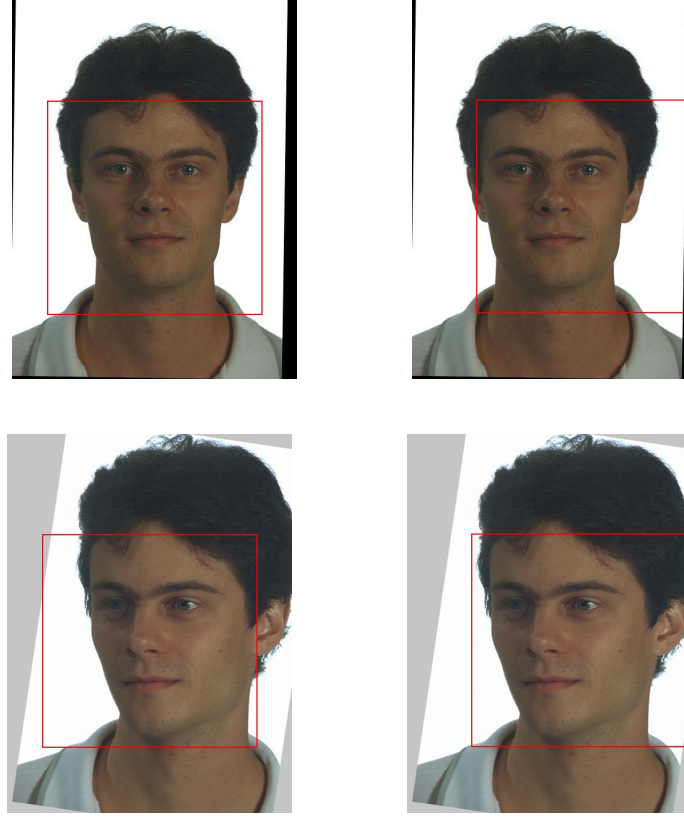


Figure 4.3: Cropping area for the training images: centered crop (left column), shifted cropping area (right column)

Table 4.1: Prior confidences for the classifier fusion of the AdaBoost based deviation from frontal pose decision.

Classifier	Prior
Left-Centered	0.39
Left-Shifted	0.37
Averaged + Left-Right	0.24

tive performance comparison is presented at several different false acceptance rate (FAR) configurations. The comparisons show the good performance of our head pose estimation approach with respect to the other vendors.



Figure 4.4: The original image and its mirrored image were overlaid and the average pixel value was taken to form a new image (left figure: frontal image, right figure: image showing pose deviation). The red squares illustrate the cropping area for AdaBoost training.

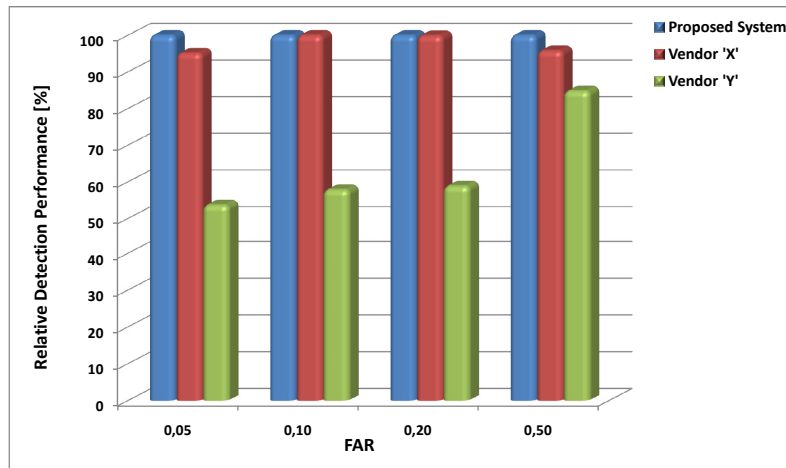


Figure 4.5: Comparison of the relative detection rates for the analysis of the deviation from frontal pose.

4.2 Person Independent Head Pose Estimation by Non-Linear Regression and Manifold Embedding

The contributions in this section deal with enhancements of existing algorithms in order to improve their pose estimation accuracy. We show how to extend the Biased Manifold Embedding [5] technique to multi-DOF pose estimation and use the histogram of oriented gradients descriptor [29] for non-linear regression [98]. In addition to that we propose a database for training a head pose estimation system (Section 4.2.3). An evaluation of all methods in an unified framework allows a direct comparison of the estimation performance of different systems (see Section 4.2.4). More details can be found in [126, 127].

We require that the HPES is person independent and can handle the large variety of faces correctly. Detecting facial features like the nose or the mouth is a challenging problem and wrongly detected features will degrade the overall performance. Therefore it is beneficial if these features need not to be detected in the input image. Only two types of methods from the overview (Chapter 4) fulfill our requirements. Both Manifold Embedding (see Section 4.2.1) as well as Non-Linear Regression methods (see Section 4.2.2) perform pose angle regression and require only a good face localization, which is given in our application. In addition to that, they are able to adapt to the variety of people found in real world scenarios automatically.

We propose an HPES similar to [98] which is sketched in Figures 4.6 and 4.7. The head pose estimation starts with a tokenized image of a human head which is presented as the input to the system. In the tokenization procedure, the eyes are detected and the input image is normalized according to their position. Therefore the roll angle can automatically be estimated by this step. A narrow region around the face is then extracted based on the position of the eyes. The yaw and pitch angles can subsequently be estimated by learning a mapping of the pixels of the region to pose angles. The remainder of this section describes two algorithms that work within such a framework.

4.2.1 Biased Manifold Embedding

The first algorithm we evaluate uses a *manifold embedding* technique [5] to estimate the pose of the human head. The main idea is to map the high dimensional input image onto a low-dimensional manifold which correctly maps pose angles while ignoring lighting, occlusions and facial differences of individuals. Pitch and yaw angles can be extracted from such a manifold by using regression techniques.

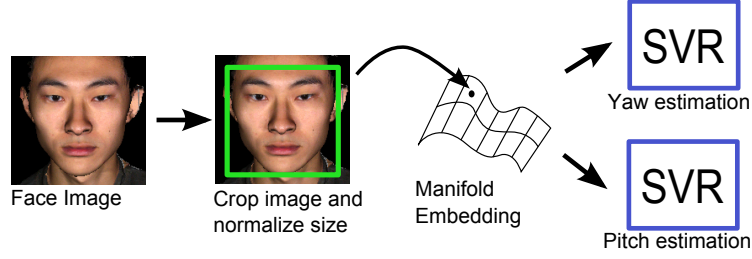


Figure 4.6: Components of a head pose estimation system using manifold embedding

We use the geometrically motivated Laplacian Eigenmaps (LE) [9] algorithm for representing high dimensional data. It allows an efficient non-linear dimensionality reduction while preserving locality properties. This makes it suitable for clustering applications such as an HPES which tries to keep similar poses within a small neighborhood on the manifold. LE require a distance measure between every tuple of data points $(\mathbf{x}_i, \mathbf{x}_j)$ from a training set such as the Euclidean distance between the i -th and j -th data point:

$$d(i, j) = \|\mathbf{x}_i - \mathbf{x}_j\| \quad (4.1)$$

As feature vectors \mathbf{x}_i we use the facial region scaled to 32×32 pixels filtered by a Laplacian of Gaussian filter in order to extract edge information. Laplacian Eigenmaps can be generated from the distances between feature vectors alone [9]. When using the unmodified distance calculation, features of the same person at different poses are likely to have a lower distance than features of different people at the same pose, which is a disadvantage in our context. Balasubramanian et al. [5] therefore enhanced manifold embedding techniques by using a *Biased Manifold Embedding* (BME) approach that incorporates pose information into the creation of the manifold in order to weight pose differences much higher than inter-person differences:

$$\tilde{d}(i, j) = \frac{|p(i, j)|}{\max_{m, n} p(m, n) - p(i, j)} \cdot d(i, j) \quad \text{with} \quad p(i, j) = |\phi_y(i) - \phi_y(j)| \quad (4.2)$$

In (4.2) this biasing term is multiplied with the Euclidean distance from (4.1). This ensures that data points with a small pose distance $p(i, j)$ for the images i and j lie close together on the resulting manifold. Balasubramanian et al. [5] use only the yaw angle $\phi_y(i)$ of the head in the image. For our system we extend the definition of the pose distance by

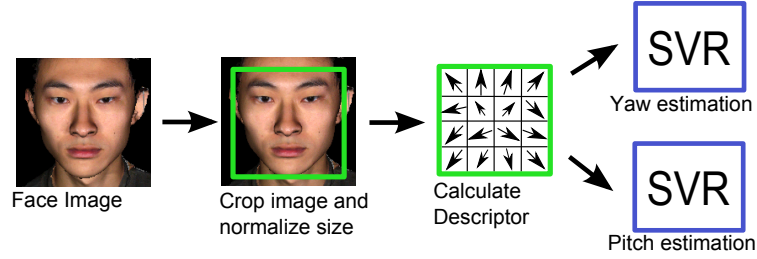


Figure 4.7: Components of a descriptor based head pose estimation system

the pitch angle $\phi_p(i)$:

$$p(i, j) = \sqrt{[\phi_y(i) - \phi_y(j)]^2 + [\phi_p(i) - \phi_p(j)]^2} \quad (4.3)$$

The complete head pose estimation process is shown in Figure 4.6. With the help of Laplacian Eigenmaps, the low dimensional manifold is created from the feature vectors of several thousand training images (see Section 4.2.3). We use two support vector regression machines to extract the yaw and pitch angles from the low dimensional manifold. As there exists no direct way to map new feature points onto an existing manifold, we learn the mapping from feature space to the manifold using a Gaussian Regression Neural Network (GRNN) [159] in order to estimate the pose of previously unseen head images.

4.2.2 Non-Linear Regression of Image Descriptors

Another approach that promises to support our requirements was initially published by Murphy-Chutorian et al. [98] who calculated a localized gradient orientation (LGO) histogram on the facial region of a normalized facial image. An LGO descriptor can be compared to a single SIFT descriptor [85] with a fixed position, orientation and scale. The system, which can be seen in Figure 4.7, extracts a scale normalized region around the face and calculated an LGO descriptor for it. Two support vector regression machines then estimate the yaw and pitch angle for the LGO descriptor vector. With this system pose angle estimation with a mean absolute pitch error of 5 degrees as well as 7 degrees in yaw angles are possible [98].

For our system we require a better performance and therefore propose several improvements. In order to reduce lighting effects, we enhance the shadow areas of the gray-level input image by a non-linear gamma normalization by transforming each pixel $I(x, y)$ to $(I(x, y))^\gamma$ with $\gamma = 0.2$. Instead of the LGO descriptor we use a Histogram of Oriented Gradients (HOG) descriptor [29] which has proven to be effective for object detection in

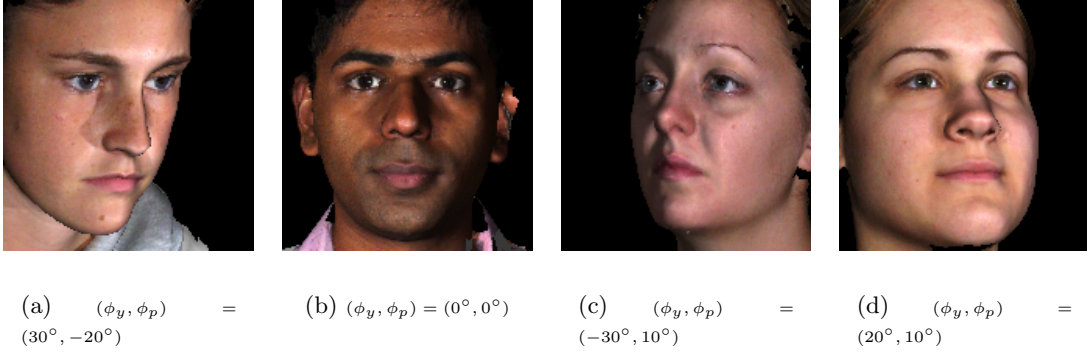


Figure 4.8: Examples of generated head poses at different yaw and pitch angles (ϕ_y, ϕ_p) and light settings

images. Recent work has pointed out that HOG descriptors are well suited for head pose classification tasks [115] but they have not been used for head pose regression so far. The major advantage of the HOG descriptor over the LGO is that the orientation of the gradients is weighted according to the gradient magnitude. In addition to that the descriptor features a spatially selective normalization which stores a single histogram multiple times but with different weights. Another important improvement deals with the scale normalization of the image. While in [98] only a scaled down facial patch is used, we follow the recommendation of [29] to use the largest scale possible without smoothing. As a result, many more gradient values are gathered in the orientation histograms of the descriptor which increases its robustness.

The effects of these enhancements on the head pose estimation accuracy will be demonstrated in Section 4.2.4. Before that we present our novel face image database which allows us to train and evaluate the systems.

4.2.3 Facial Image Database

For training an image based head pose estimation system, a large database of facial images in different poses and with multiple people is needed. A variety of databases are mentioned in the literature [12, 98] but there is either not enough variation in the images or the database is not open to the public. This is why we propose a method to create a face image database with a high flexibility at a very low cost by rendering 3D models from existing laser scans. The Facial Expression database from the Binghampton University [155] contains fully textured 3D heads of 100 subjects. We generate multiple facial images from each head model by rotating it in 3D space with yaw within $\pm 50^\circ$ and pitch

angles within $\pm 30^\circ$. The background of each rendered image is set to an arbitrary image in order to simulate a non-uniform background. In addition to that, we rotate a variable light source around the head in order to generate a variety of lighting situations. Such training data allows the system to become more robust and person independent. In all generated images the mid-point of the eyes is kept at a constant position which allows us to eliminate the face detection prior to head pose estimation. Therefore our training images are not influenced by the performance of the head localization step. Some of the synthetic images can be seen in Figure 4.8.

4.2.4 Experimental Results

We performed an experimental evaluation on our artificial database (see Section 4.2.3) as well as on the publicly available FacePix database [12]. Three different implementations of a head pose estimation system were evaluated: Biased Manifold Embedding [5], Local Gradient Orientations mapped by an SVR [98] and our Histograms of Oriented Gradients based improvement described in this section.

Training of the systems was performed by using a subset of head images from persons in our database while keeping the remaining persons for testing. The optimal parameters for the descriptors and machine learning algorithms were found by an exhaustive search over the parameter space. As the performance measure we used the mean absolute pose angle estimation error. This measure is often used in the literature and therefore allows comparison with other publications (e.g. [100]):

$$E_y = \frac{1}{N} \sum_{i=1}^N |\phi_y(i) - P_y(\mathbf{D}_i)| \quad E_p = \frac{1}{N} \sum_{i=1}^N |\phi_p(i) - P_p(\mathbf{D}_i)| \quad (4.4)$$

It is defined as the absolute difference between the ground-truth pose angle of the i -th image (e.g. $\phi_y(i)$) and the predicted angle from the descriptor \mathbf{D}_i of the image. The mean over all N images in the test-set yields the final value for either yaw or pitch estimation errors.

In Table 4.2 the performance of the systems on yaw and pitch angle estimation is compared using the mean absolute error. It can clearly be seen that the HOG approach outperforms the other methods in both pitch and yaw angle estimation performance. The biased manifold embedding approach shows the worst results in our evaluation. This finding is somewhat contrary to what is stated in [5] where mean absolute errors of around 2° are promised. The reason for this is that the original algorithm is trained and tested

Method	E_y our DB	E_p our DB	E_y FacePix
Biased Manifold Embedding	5.6°	6.5°	14.0°
LGO / SVR	4.5°	4.4°	7.5°
HOG / SVR	3.6°	3.1°	4.0°

Table 4.2: Comparison of the Biased Manifold Embedding, LGO and HOG approach in terms of mean absolute pose angle estimation error when trained and tested on our database and additionally tested on the FacePix database [12]

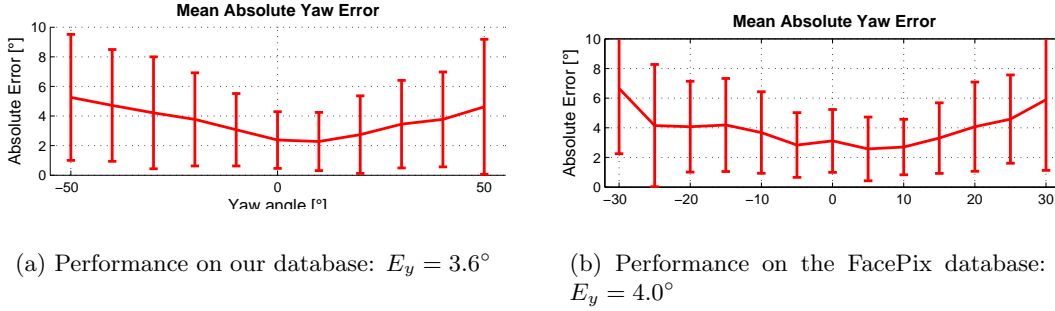


Figure 4.9: Yaw angle estimation performance of the HOG descriptor with SVR learning for different yaw angles (mean and standard deviation)

on the FacePix database which only contains rotations around a single axis (left/right). When used in a multi-dimensional pose estimation system or heterogeneous datasets, this performance degrades due to ambiguities between neighboring poses.

A more detailed evaluation of the HOG descriptor based system in Figure 4.9(a) shows that the yaw angle estimation performance has a U-shaped curve for different yaw angles with a minimum at the 0° yaw angle. This means that yaw angles are estimated more accurately in frontal pose head images. Such behavior is beneficial in applications such as ours where we want to estimate the head pose in near-frontal images. A similar evaluation in Figure 4.9(b) shows a good performance of the HOG based system with an average yaw estimation error of 4.0° on the FacePix database [12] while still being trained on our artificial database.

As the ICAO standard [58] allows only frontal head poses for passport images, we evaluate the performance of the classification of head images into frontal and non-frontal poses as well. We do this by thresholding the estimated yaw and pitch angles for a given head pose image. In Figure 4.10 we show the Receiver Operating Characteristics (ROC) as true and false positive rates for different threshold values on non-training images from our generated database. At a threshold of slightly more than five degree we are able to classify

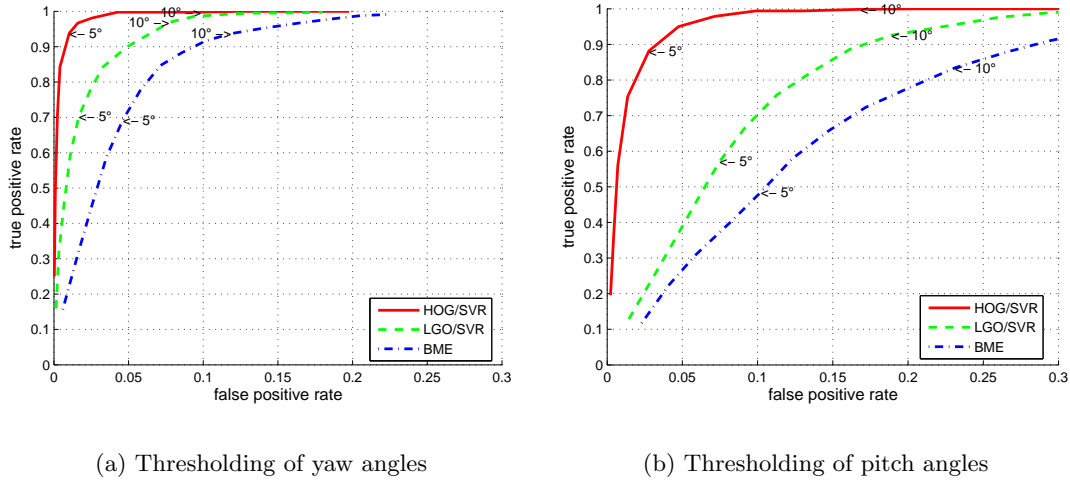


Figure 4.10: Evaluation of the frontal-pose classification performance of the three system using ROC plots

over 90% of all faces correctly with less than 5% false positives when using the HOG based HPES, which is due to the low estimation error in frontal poses. Other methods do not show such a good performance for frontal pose classification.

4.2.5 Summary

In this section we presented a system for head pose estimation from monocular still images. We estimated continuous pose angles from either localized gradient orientation (LGO) histogram descriptors, histogram of oriented gradients (HOG) descriptors or manifold embedded pixel values. The main contributions were the extension of the biased manifold embedding approach [5] towards pose estimation in two DOF and use of the HOG descriptor [29] for head pose estimation. While the HOG descriptor has become quite common in object detection, we showed that its advantages hold in regression tasks as well. The biased manifold embedding approach did not show convincing results in our experiments when used for the estimation of more than one pose angle.

For training the HPES a novel database was generated from 3D head models, also background clutter and lighting variations were simulated. We showed that a system trained on our database also performed well on different databases.

4.3 3D-MAM: 3D Morphable Appearance Model for Efficient Fine Head Pose Estimation from Still Images

According to the ICAO standard, persons are required to show a frontal head pose which means that the head rotation must not deviate more than ± 5 degrees in any direction from frontal. Thus, we present a novel 3D head pose estimation approach, which utilizes the flexibility and expressibility of a dense generative 3D facial model in combination with a very fast fitting algorithm. The efficiency of the head pose estimation is obtained by a 2D synthesis of the facial input image. This optimization procedure drives the appearance and pose of the 3D facial model. We evaluate our approach on two publicly available databases (FacePix and USF HumanID) and compare our method to the 3D morphable model and other state of the art approaches in terms of accuracy and speed.

We focus on flexible model based head pose estimation to overcome problems of other approaches of being not adaptive to unseen facial images. Our proposed approach is related to the well known Active Appearance Model (AAM) of Cootes et al. [22]. The AAM is a widely used method for model based vision showing excellent results in a variety of applications. As a generative model it describes the statistical variation in shape and texture of a training set representing an object. AAM model fitting is performed in a gradient descent optimization scheme, where the cost function is defined as the L2 norm of the intensity differences. For optimization, the Jacobian is often approximated either by a regression to learn the dependency between model parameter updates and intensity differences [22] or alternatively by a canonical correlation analysis [31]. The fitting procedure is very fast, but one major drawback is its non-robustness to viewpoint changes. In the case of facial images, the AAM fitting is not appropriate for adapting to faces which exhibit pose variations. Cootes et al. [24] extend their AAM approach for multi pose fitting by combining a small number of 2D AAM models.

Blanz and Vetter [15, 16] overcome the drawbacks of the 2D AAM by creating a 3D Morphable Model (3DMM). The 3DMM is a statistical model of shape and texture based on data acquired from a laser scanner. The approach shows amazing image synthesis results but the fitting procedure is computationally very expensive. One attempt to fit a 3DMM more efficiently was proposed in [116], but the fitting of one facial image still takes several minutes.

In [25] an extension to the classical AAM approach is proposed. They build a 3D anthropometric muscle based active appearance model using a generic 3D face shape. They adapt the 3D shape model so that the projected 3D vertices best fit to a facial 2D

image. Using several adaptations to different facial images, these obtained shapes can be taken to create a shape eigenspace using PCA. According to the foregoing shape adaption the texture of the 2D facial images is warped back onto the 3D model, i.e., they get several textures to create a texture eigenspace. This generation of training data is a cumbersome work. The model fitting is similar to the original AAM fitting procedure. They apply their approach for head tracking and facial expression recovery [26].

Xiao et al. [151] propose a real time combined 2D+3D AAM to fit 3D shapes to images. They also investigate, that the 2D AAM could generate illegal model instances, which do not have a physical counterpart. They show how to constrain an AAM by incorporating 3D shape information so that the AAM can only generate valid model instances. Their work focuses on tracking applications and experiments show excellent performance, however, in their setup the generative model is always built from the same person that is tracked later. Chen and Wang [21] describe a similar model for human-robot interaction.

In the domain of medical image analysis, 3D Active Appearance Models are used with great success for segmentation [7, 94, 125] and modeling of shape or pathological variations of a population. These approaches are specifically targeted to 3D volumetric data, where the notion of efficiency becomes even more important due to the increased dimensionality.

Due to the two major drawbacks of model based vision, the non-robustness to viewpoint changes and the inefficient fitting procedure, we present a novel 3D Morphable Appearance Model (3D-MAM) for head pose estimation, which utilizes the flexibility and expressibility of a dense generative 3D facial model in combination with a very fast fitting algorithm. The efficiency of the head pose estimation is reached by a 2D synthesis of the facial input image. This optimization procedure drives the appearance and pose of the 3D facial model. In contrast to many other approaches we are specifically interested in the more difficult task of head pose estimation from still images, instead of tracking faces in image sequences. Much effort is undertaken to build a fair evaluation scheme for our approach. That is, the data for building and training of our 3D-MAM was totally independent of the datasets used for the evaluations.

This section is structured as follows: In Section 4.3.1 we introduce and discuss our 3D-MAM approach in terms of model building and model fitting. In Section 4.3.2 we present our results by evaluating the head pose estimation accuracy and speed of our approach on two publicly available databases (FacePix and USF HumanID) and compare our method to state of the art approaches. Finally, we discuss our findings and summarize our work in Section 4.3.3.

4.3.1 3D Morphable Appearance Model

We build a generative 3D Morphable Appearance Model (3D-MAM) based on registered laser scans of human heads. The advantage of a dense 3D model is its flexibility to express generic faces and the ability to adapt to non-rigid deformations exhibited by faces. Only with a dense model, normal vectors of the surface can be computed and therefore depth can be estimated correctly. This resembles human perception of 3D objects. A human is only able to estimate depth correctly in the presence of shadowed surfaces.

To overcome the slow fitting performance exhibited by many approaches using dense 3D models, we perform our fitting step in 2D (Section 4.3.1.2). To sum up, we combine the advantages of dense 3D models and the very efficient fitting speed gained in the 2D domain.

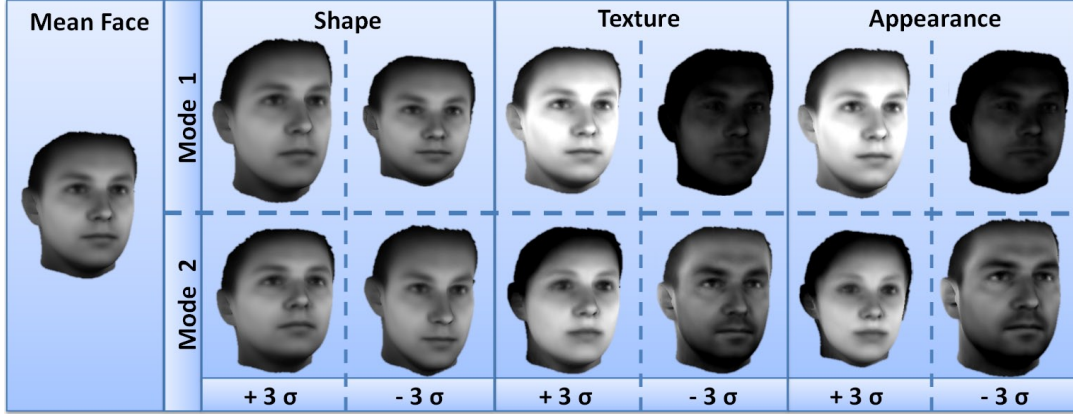


Figure 4.11: 3D Morphable Appearance Model. Effect of varying the shape, texture and appearance parameters of the first and second mode by ± 3 standard deviations.

4.3.1.1 3D Model

We utilize laser scans of human heads and register them using a nonrigid Iterative Closest Point (ICP) algorithm [3]. The registered 3D laser scans form the basis for building a parameterized generative 3D model, which will later be used for head pose estimation.

3D Head Laser-Scans The 3D facial model is built from 350 facial laser scans. The scans were acquired by a *Cyberware*TM laser scanner, which captures the 3D information (vertices) in cylinder coordinates with radius $r(h, \phi)$, 512 equally-spaced angles ϕ and 512 equally-spaced vertical steps h . Additionally, the RGB-color information $R(h, \phi), G(h, \phi), B(h, \phi)$ is recorded for each vertex.

The obtained 3D database consists of 350 different subjects exhibiting a wide variability in race, gender and age. Most of the subjects show neutral facial expression. The raw facial scans have to be post-processed to remove certain parts of the scans, e.g., hair and shoulders. Often those areas cannot be captured very well, because of the fine structure of the hair or due to self occlusions. More specifically, the scans are cut vertically behind the ears and cut horizontally to remove the hair and shoulders. Additionally, laser scanning artifacts, like holes or spikes, are removed manually.

We reduce the amount of vertices from about 100,000 to 10,000 for the purpose of decreasing the computational effort in the model fitting procedure, see Section 4.3.1.2. This simplification of the 3D data at regions with little details is performed by a structure preserving surface simplification approach [45].

To build a generative model (Section 4.3.1.1), the individual laser scans have to be non-rigidly registered. Blanz and Vetter [15] register their data using a modified gradient-based optical flow algorithm. We use the more sophisticated *optimal step nonrigid ICP* method [3] to establish correspondence between a pair of 3D scans. This method's runtime is slower compared to optical flow, but yields more robust registration results, e.g., filling of holes due to missing data.

Model Building We create a statistical model of shape, texture and appearance similar to [22] with the difference of using 3D laser scanner data, instead of annotated 2D images. The laser scanner data have to be registered (Section 4.3.1.1) to allow the construction of a generative model utilizing Principal Component Analysis (PCA).

The registered 3D shapes are composed of the 3D positions of the vertices, and the texture consists of the intensity values of the vertices. Taking N training shape and texture tuples with sample mean $\bar{\mathbf{s}}$ and $\bar{\mathbf{t}}$ correspondingly, they are used to build statistical models of shape and texture by using PCA

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{U}_s \mathbf{p}_s \quad (4.5)$$

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{U}_t \mathbf{p}_t \quad (4.6)$$

Here \mathbf{U}_s and \mathbf{U}_t are shape- and texture eigenvectors, which describe the modes of variation derived from the training set. By adjusting the parameters \mathbf{p}_s and \mathbf{p}_t , new instances of shape \mathbf{s} and texture \mathbf{t} can be generated.

To remove correlations between shape and texture variations, we apply a further PCA

to the data. The shape- and texture eigenspaces are coupled through

$$\begin{pmatrix} \mathbf{W}_s \mathbf{p}_s \\ \mathbf{p}_t \end{pmatrix} = \mathbf{U}_c \mathbf{c} \quad (4.7)$$

to get the statistical model of appearance (combined model), where \mathbf{W}_s is a diagonal scaling matrix to compensate for the different measure units of shape and texture. \mathbf{W}_s is defined as the ratio of the total intensity variation to the total shape variation [22]. This appearance model is controlled by parameter \mathbf{c} to obtain new instances of facial shape and texture

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{U}_s \mathbf{W}_s^{-1} \mathbf{U}_{c,s} \mathbf{c} \quad (4.8)$$

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{U}_t \mathbf{U}_{c,t} \mathbf{c} \quad (4.9)$$

$$\mathbf{U}_c = \begin{pmatrix} \mathbf{U}_{c,s} \\ \mathbf{U}_{c,t} \end{pmatrix} \quad (4.10)$$

Figure 4.11 shows the effect of varying the shape, texture and appearance parameters of the first and second mode by ± 3 standard deviations obtained from the training set.

The pose of the appearance model in 3D can be altered by six degrees of freedom (DOF), i.e., three angles of rotation and the three directions of translation. We map the 3D points to the 2D image coordinates by a weak perspective projection. That is, the rendering of the 3D model to the image plane is given by the two rotation angles θ_{pitch} and θ_{yaw} and the two translations u_x and u_y in the image plane. For linearity, the *scaling* and the *roll* angle is represented as $sr_x = (scale \cos \theta_{roll} - 1)$ and $sr_y = scale \sin \theta_{roll}$. The concatenation of those single parameters yields the pose parameter vector $\mathbf{p}_{pose} = (sr_x, sr_y, \theta_{pitch}, \theta_{yaw}, u_x, u_y)$.

4.3.1.2 Model Fitting

The model is fitted iteratively in an analysis-by-synthesis approach, see Figure 4.13. A direct optimization of the appearance model parameters \mathbf{c} and pose parameters \mathbf{p}_{pose} is computationally not feasible for real time applications. Hence, we precompute a parameter update matrix [22], which will be used to incrementally update the parameters $\mathbf{p} = (\mathbf{c}, \mathbf{p}_{pose})$ in the fitting stage. The patch used for synthesizing a new input image is restricted to an area of the head, where most of the vertices are visible for slight pose variations (Figure 4.12).

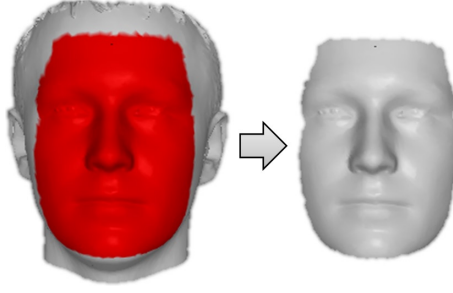


Figure 4.12: Patch extracted from the whole head used for synthesizing a new input image.

Starting from the 3D mean shape, we project the positions of the vertices from the 3D patch to the 2D input image using a weak perspective projection. The texture from the input image underlying the projected points in 2D is then warped to a shape free (mean shape) representation. Simultaneously, the texture of the model patch is rendered also to the same shape free representation. Now, the texture from the input image \mathbf{t}_s and the rendered texture from the model patch \mathbf{t}_m (both in the same shape free representation) can be subtracted to get a residual image

$$\mathbf{r}(\mathbf{p}) = \mathbf{t}_s - \mathbf{t}_m. \quad (4.11)$$

A first order Taylor expansion of (4.11) gives

$$\mathbf{r}(\mathbf{p} + \delta\mathbf{p}) = \mathbf{r}(\mathbf{p}) + \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \delta\mathbf{p} \quad (4.12)$$

In the fitting stage, $\|\mathbf{r}(\mathbf{p} + \delta\mathbf{p})\|^2$ is minimized by computing

$$\delta\mathbf{p} = -\mathbf{R}\mathbf{r}(\mathbf{p}) \quad \text{where} \quad \mathbf{R} = \left(\frac{\partial \mathbf{r}^T}{\partial \mathbf{p}} \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \right)^{-1} \frac{\partial \mathbf{r}^T}{\partial \mathbf{p}} \quad (4.13)$$

In a direct optimization scheme, the Jacobian matrix $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ has to be recomputed in every iteration step yielding poor runtime performance. Hence, the parameter update matrix \mathbf{R} is assumed to be fixed and can therefore be precomputed by numeric differentiation. The numeric differentiation is accomplished through a perturbation scheme, i.e., each parameter is displaced from a known optimal value. More details can be found in [22].

In the fitting stage, texture residuals are computed in the same way as in the training stage of the parameter update matrix \mathbf{R} . This residual in combination with the update matrix gives the parameter update $\delta\mathbf{p}$ for driving the parameters \mathbf{p} of the 3D model. That is, the appearance and pose of the 3D model is iteratively fitted to the 2D input image.

The whole fitting procedure is illustrated in Figure 4.13.

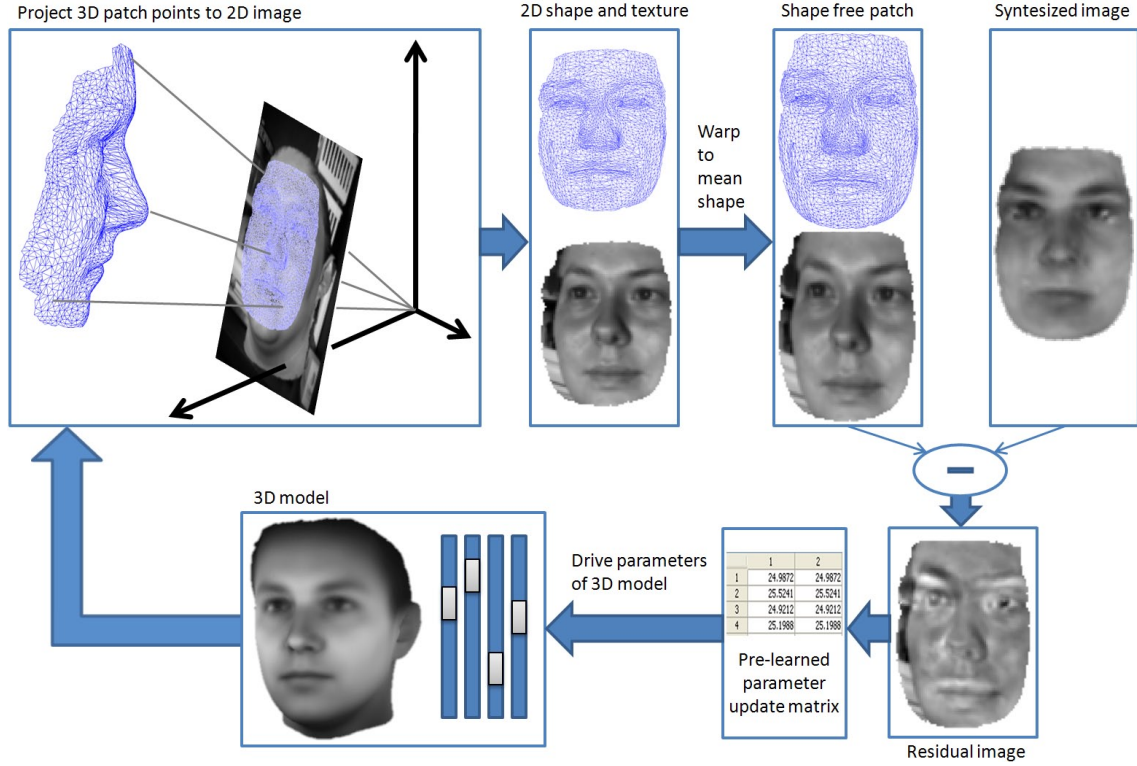


Figure 4.13: Fitting workflow

4.3.2 Experimental Results

We build a 3D-MAM (Section 4.3.1.1) and keep 90% of the eigenvalue energy spectrum for the shape, 85% for the texture and 90% of the appearance variation to represent our compact model. We precompute the parameter update matrix (Section 4.3.1.2) with a resolution of the fitting patch of 60x80 pixels.

The head pose estimation is evaluated on two different publicly available datasets (USF Human ID 3D face database and FacePix). Those data sets are independent of the data used for model building.

The USF Human ID 3D face database [138], [15] consists of 136 individuals, which are recorded by a *Cyberware*TM laser scanner. Each facial model is composed of more than 90,000 vertices and 180,000 triangles. Images of the individuals can be rendered in arbitrary pose. Those rendered images with arbitrary textured background added are used as test images for our approach. First, we evaluated the head pose estimation capability

of our approach using the rendered images of the first 50 individuals in the database by altering only the yaw angles from -16° to $+16^\circ$ in steps of 4° while fixing the roll- and pitch angle of the rendered test images to zero. The 3D-MAM's 2D starting position is roughly initialized manually. In the future, this initialization will be done by an automatic face- and facial feature detection stage. Figure 4.14a presents the mean and standard deviation of the absolute angular error for the single yaw rotations. Table 4.3a summarizes the error measures for the whole range of rotations.

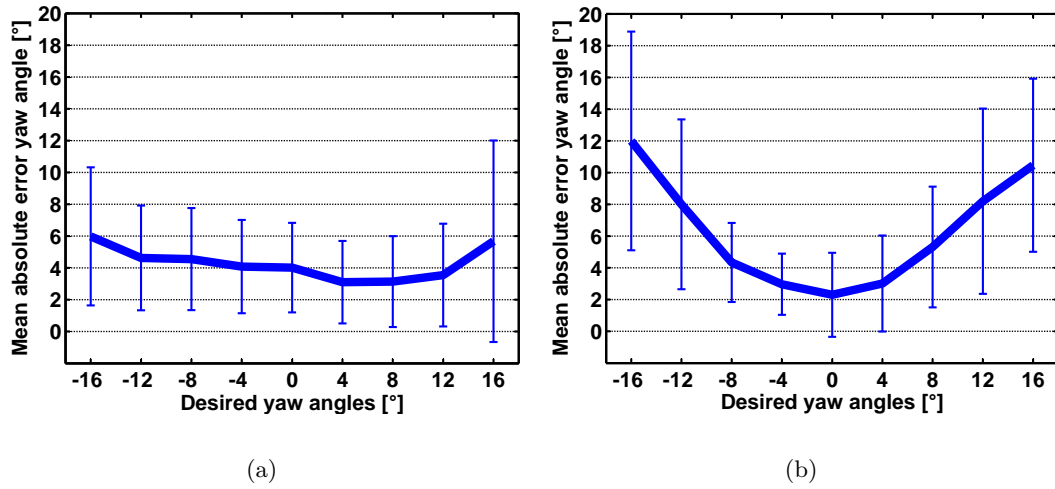


Figure 4.14: Mean and standard deviation of the absolute yaw angular error for the (a) USF Human ID 3D face database and (b) FacePix database.

Table 4.3: Mean, standard deviation, median, upper- and lower quartile of the absolute yaw angular error by only altering the yaw angle for the (a) USF Human ID 3D face database and (b) FacePix database. The results are compared to the 3DMM [15].

(a)

USF Human ID	Absolute Error [°]				
	mean	std	median	Q ₂₅	Q ₇₅
Our Approach	4.30	3.51	3.35	1.71	6.08

(b)

FacePix	Absolute Error [°]				
	mean	std	median	Q ₂₅	Q ₇₅
Our Approach	6.29	4.15	4.62	2.30	8.57
3DMM	4.89	3.15	4.48	2.15	7.06

We extend the previous experiment by altering the yaw- and pitch angle by $[-16^\circ \ 0^\circ \ +16^\circ]$ and $[-8^\circ \ 0^\circ \ +8^\circ]$ correspondingly. These nine angle combinations and the 50 individuals per combination yields 450 3D-MAM fitting runs. The mean absolute angular error of the angle combinations is shown in Figure 4.15(a,b). The error measures are summarized in Table 4.4a. We compare our pose estimation results with the well known 3D-Morphable Model [15]. We build a 3DMM based on the same laser scanner data as used for our 3D-MAM (Section 4.3.1.1). To speed up the 3DMM fitting procedure, we use only the first 10 shape- and texture modes, because we want to estimate head pose and do not want to synthesize the test image in every detail. The results for the 3DMM are shown in Figure 4.15(c,d) and summarized in Table 4.4a. The 3DMM exhibits slightly better head pose estimation results at the cost of a much higher runtime per facial fit, see Table 4.4b.

Table 4.4: Evaluations for the USF Human ID 3D face database. (a) Mean, standard deviation, median, upper- and lower quartile of the absolute yaw and pitch angular error by altering the yaw- and pitch angle. (b) Average runtime* per facial fit. The results are compared to the 3DMM [15].

(a)

USF Human ID	Absolute Error [°]									
	Yaw					Pitch				
	mean	std	median	Q _{.25}	Q _{.75}	mean	std	median	Q _{.25}	Q _{.75}
Our Approach	5.78	4.22	4.86	2.57	7.66	5.89	4.68	4.76	2.32	8.23
3DMM	3.90	3.31	2.81	1.55	5.21	5.14	3.66	4.08	2.11	7.55

(b)

	Average Runtime [s]
Our Approach	3.2
3DMM	33.5

The second database, CUBiC FacePix(30) database [12, 84], consists of 30 individuals. For each individual, three sets of images are available. The first set contains images taken from the individuals' right to left (only yaw angle is annotated), in one degree increments. The second- and third set is targeted to non-uniform lighting experiments. We are specifically interested in the first set for our pose estimation experiments. We take those images annotated by -16° to $+16^\circ$ in steps of 4° . The mean and standard deviation of the absolute angular error is shown in Figure 4.14b. Table 4.3b summarizes the error measures for the whole range of rotations and compares the results to the 3DMM

approach. In [6], they also conducted several experiments on the FacePix(30) database using manifold embedding methods. They show better results, ranging from a mean absolute error of 1.44° to 10.41° , but they performed the training and testing on the same database in a cross-validation scheme, which leaves serious doubts about the general applicability of their method on unseen data. Second, they are limited to only estimate the yaw angle of a given test image.

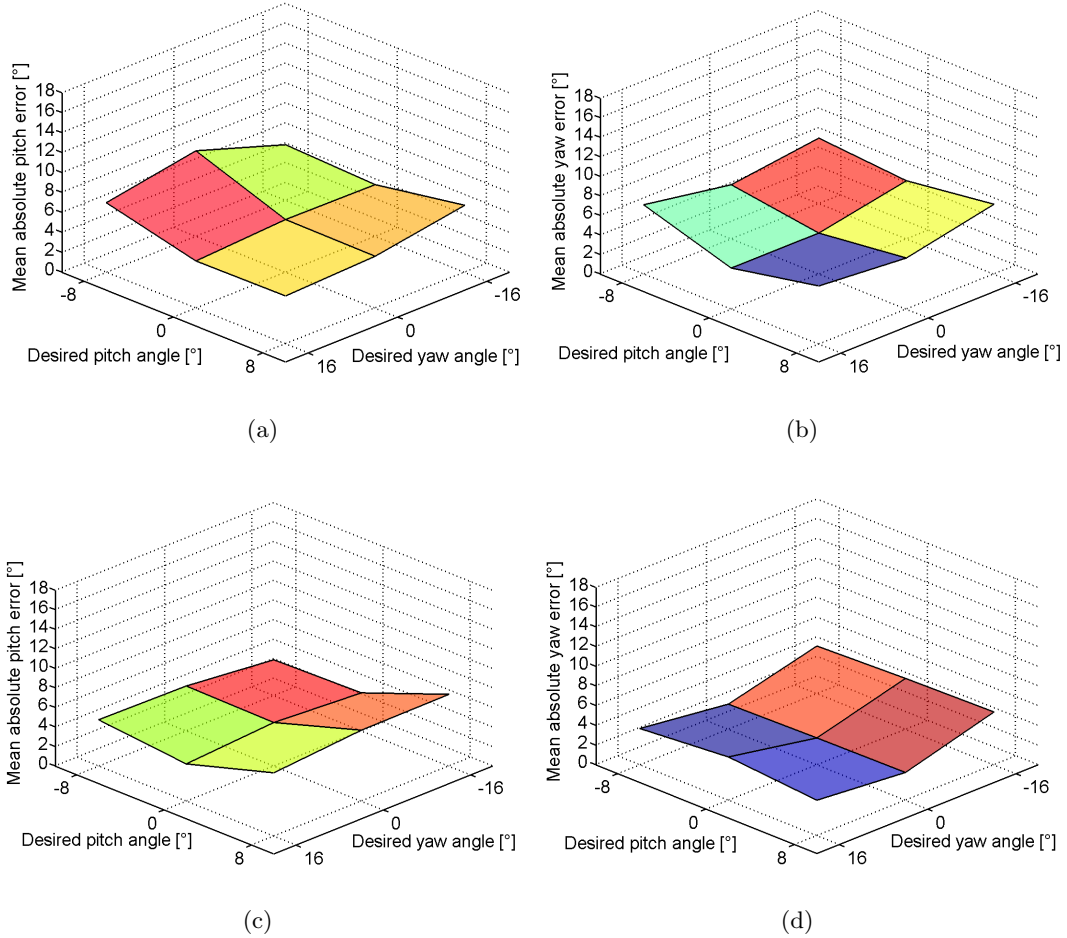


Figure 4.15: Mean absolute error for pitch- and yaw angle on the USF Human ID 3D face database by altering pitch- and yaw angles for the generation of the test images. (a,b) Our approach (c,d) 3DMM

Our model fitting strategy is similar to the AAM approach [22], leading to an excellent runtime performance. The average runtime* for our approach is 3.2s at an average number

*The runtimes are measured in *MATLAB*TM using an Intel Core 2 Duo processor running at 2.4GHz. The resolution of the images is 60x80 pixels.

of iterations of 14. We have a comparable implementation of an AAM in C++, which takes about 15ms per facial fit. If we add about 5ms per iteration for a rendering of a facial image using OpenGL, and taking the average number of iterations into account, we would get an estimated average runtime of 85ms per facial fit with an implementation of 3D-MAM in C++. This runtime would enable the usage of our approach for real-time head pose estimation.

Figure 4.16 shows frames from a 3D-MAM facial fit starting with the mean model. During fitting the patch synthesizes the face and adjusts the 3D model in appearance and pose.

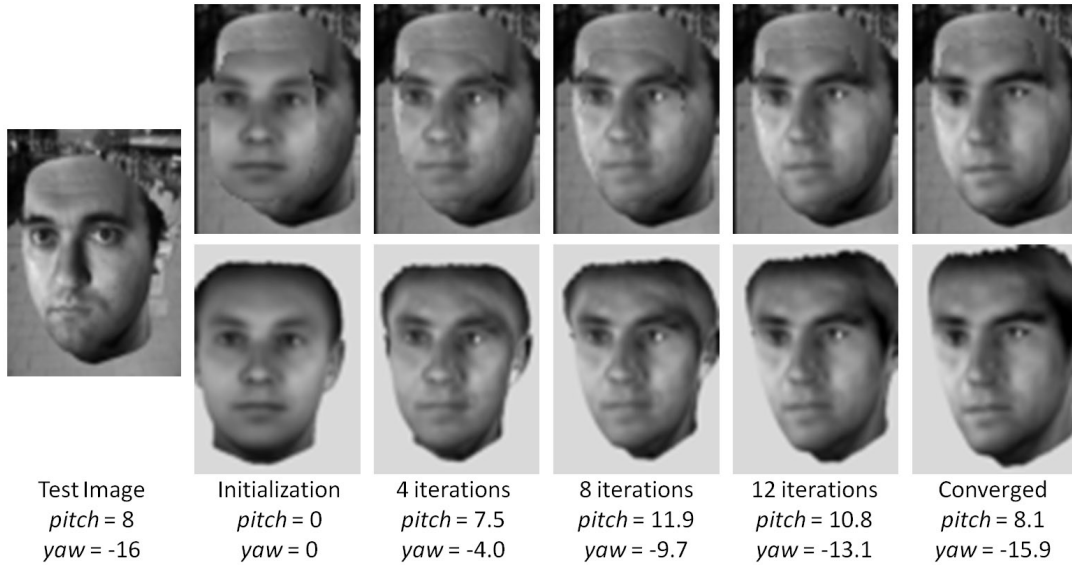


Figure 4.16: Analysis-by-Synthesis fitting example. The lower row shows the adjustment of the 3D model. In the upper row, the corresponding fitting patch is illustrated. This model fitting converges after 17 iterations.

4.3.3 Summary

Two major shortcomings of existing model based head pose estimation approaches, the non-robustness to viewpoint changes and the inefficient fitting procedure, motivated us to generate a 3D-MAM. It utilizes the flexibility of a dense 3D facial model combined with a very fast fitting algorithm in 2D. In the experiments, we show the applicability of our approach for head pose estimation on two publicly available databases (USF HumanID and FacePix). We compare our results to state of the art head pose estimation algorithms in terms of accuracy and speed.

Occlusion Handling

The ICAO standard prohibits photographs showing occlusions, thus there is the need to detect occluded images automatically. In Section 5.1 we present a novel algorithm for occlusion detection and evaluate its performance on several databases. First, we use the publicly available AR faces database which contains many occluded face image samples. We show a straight-forward algorithm based on color space techniques which gives a very high performance on this database. We conclude that the AR faces database is too simple to evaluate occlusions and propose our own, more complex database, which includes, e.g., hands or arbitrary objects covering the face. Finally we extend our first algorithm by an Active Shape Model in combination with a PCA reconstruction verification. We show how our novel occlusion detection algorithm outperforms the simple approach on our more complex database.

In the case of an occlusion, we cannot only detect an occlusion, but we also can improve the fitting quality of an Active Appearance Model (AAM), see Section 5.2. The AAM is a widely used method for model based vision showing excellent results. But one major drawback is that the method is not robust against occlusions. Thus, if parts of the image are occluded the method converges to local minima and the obtained results are unreliable. To overcome this problem we propose a robust AAM fitting strategy. The main idea is to apply a robust PCA model to reconstruct the missing feature information and to use the obtained image as input for the standard AAM fitting process. Since existing methods for robust PCA reconstruction are computationally too expensive for real-time processing we developed a more efficient method: Fast-Robust PCA (FR-PCA), see Section 5.3. In fact, by using our FR-PCA the computational effort is drastically reduced. Moreover,

more accurate reconstructions are obtained. In the experiments, we evaluated both, the fast robust PCA model on the publicly available ALOI database and the whole robust AAM fitting chain on facial images. The results clearly show the benefits of our approach in terms of accuracy and speed when processing disturbed data (i.e., images containing occlusions).

5.1 Occlusion Detection

Starting with the standardized coordinate frame (Section 2), one can derive criteria to define images with and without occlusions. In this section we concentrate on occlusions due to extraordinary glasses and objects covering parts of the face (hands, hair, or other objects). See Figure 5.3 and Figure 5.5 for some examples.

The problem of occlusions in the context of face recognition has recently been studied in [37], where the authors show different amounts of degradation in recognition performance depending on the location of the facial occlusion. A number of techniques have emerged to make the recognition algorithm itself robust to occlusions. Early work in this direction has been proposed by Leonardis and Bischof [75] who showed how to handle occlusions in an eigenface [137] framework. Their key idea was to extract eigenspace coefficients by a robust hypothesize-and-test paradigm using subsets of image points instead of computing the coefficients by projecting the data onto the eigenimages. Li et al. [77] presented a local non-negative matrix factorization (LNMF) to learn spatially localized part based subspace representations from visual patterns. Their use of localization constraints showed good performance on the AR face database. Extending this work, Oh et al. [104] proposed a selective LNMF technique with a partial occlusion detection step on a number of disjoint image patches. These patches are represented by a PCA to obtain corresponding occlusion-free patches, followed by the LNMF procedure used exclusively on the bases of the occlusion-free image patches. A different direction was pursued by Martinez [89] who described a probabilistic approach that compensates for imprecisely localized, partially occluded faces under different facial expressions. He divides the face into a number of local regions and matches them to a single prototype by a probabilistic scheme. He demonstrates robustness in the presence of occlusion of $1/6$ to $1/3$ of the facial area at the cost of only a slight decrease in accuracy. All of these presented approaches have in common that their goal is to perform face recognition in the presence of occlusions. However, in our task we explicitly want to detect occlusions to sort out unsuitable images for a subsequent recognition step. This is in accordance with the ICAO specification which

prohibits occluded facial images.

A tightly related topic is face hallucination which was made popular in work by Baker and Kanade [4]. Here occluded parts of a face are recovered by using generative face models. Different terms describing this area of research are face recovery and regeneration or face image inpainting. Some examples of recent work are presented in [83, 156]. Our face images are used for machine-readable travel documents, so we do not want to modify given occluded images. Therefore we do not focus further on this research direction.

An obvious choice for an occlusion detection algorithm is the widely used Active Appearance Model (AAM) [22]. Here the strategy is to use the generative AAM model fitting algorithm starting with a suitable model initialization on a face portrait image. By fitting the model to the occluded image one could derive a quality measure (e.g., the final sum-of-squared differences) to make a decision if an occlusion is present or not. Here, the main problem is that the original AAM model formulation is not very robust to occlusions. Some extensions of the AAM model in the presence of occlusions have been presented in the literature [47, 157], however their holistic approach poses a basic difficulty during model fitting, since the quality measure driving the fitting optimization always is influenced by the occluded part to a certain degree and the non-convex optimization is prone to get stuck in local minima. Due to their popularity and widespread availability we will show where this class of algorithms tends to fail in our occlusion detection task.

In this section we propose a novel system to automatically detect occlusions from tokenized facial images in Section 5.1.1. This is an important pre-processing step for the training of face recognition/verification, but could also be used for the testing step. In Section 5.1.1.1 we start with a straight-forward occlusion detection method based on color space techniques and perform occlusion detection experiments on the publicly available AR database [90]. We show why this database is not sufficient to evaluate an algorithm for facial occlusion detection, and we present our own more challenging database which we created specifically for this task. We extend our first algorithm and include an Active Shape Model (ASM) [23] approach followed by a PCA based verification step described in Section 5.1.1.2. This second method is able to solve the occlusion detection problem on our own more difficult database. Finally, we discuss and summarize our findings in Section 5.1.2.

5.1.1 Methodology

We start with a simple and straight-forward approach for occlusion detection in Section 5.1.1.1 which we refer to as *Method 1*. It is based on automatic color correction techniques and on the HSV color space. It turns out that this method is already well suited and sufficient to detect occlusions on the publicly available AR face database [90].

We created our own collection of images which extends the variations exhibited by the AR database in terms of further illumination conditions and types of occlusions. We show that the simple Method 1 does not perform very well on this more challenging database. Hence, we exploit our findings of our color experiments of Method 1 and extend this first method by an Active Shape Model [93] in combination with a projection of facial parts to separate Principal Component (PCA) subspaces. We refer to this extension as *Method 2* explained in detail in Section 5.1.1.2.

In our proposed system we make use of input images in the tokenized coordinate frame according to the ICAO specification. In order to be able to analyze arbitrary facial images we have to transform them first into this coordinate frame based on eye locations. For this purpose we use a robust face and facial component detection stage followed by a probabilistic voting scheme for the most probable face and eye position, Section 2. The tokenized image is finally derived by warping the input image according to the eye locations.

5.1.1.1 Method 1

Our first approach is based on automatic color correction and on the H-channel of the HSV color space. Before transforming the image into the HSV color space an automatic color correction is applied. It reduces the effects of global illumination and could also be referred to as automatic white balancing based on color temperatures.

Our automatic color correction algorithm assumes that the average surface color in a scene is gray. This means that the shift from gray of the measured averages on the three channels corresponds to the color of the illuminant. Three scaling coefficients, one for each color channel, are therefore set to compensate this shift [11, 18]. Every RGB-pixel value is adjusted according to

$$\begin{pmatrix} r_{adj} \\ g_{adj} \\ b_{adj} \end{pmatrix} = \begin{pmatrix} \bar{Y}/\bar{R} & 0 & 0 \\ 0 & \bar{Y}/\bar{G} & 0 \\ 0 & 0 & \bar{Y}/\bar{B} \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix}, \quad (5.1)$$

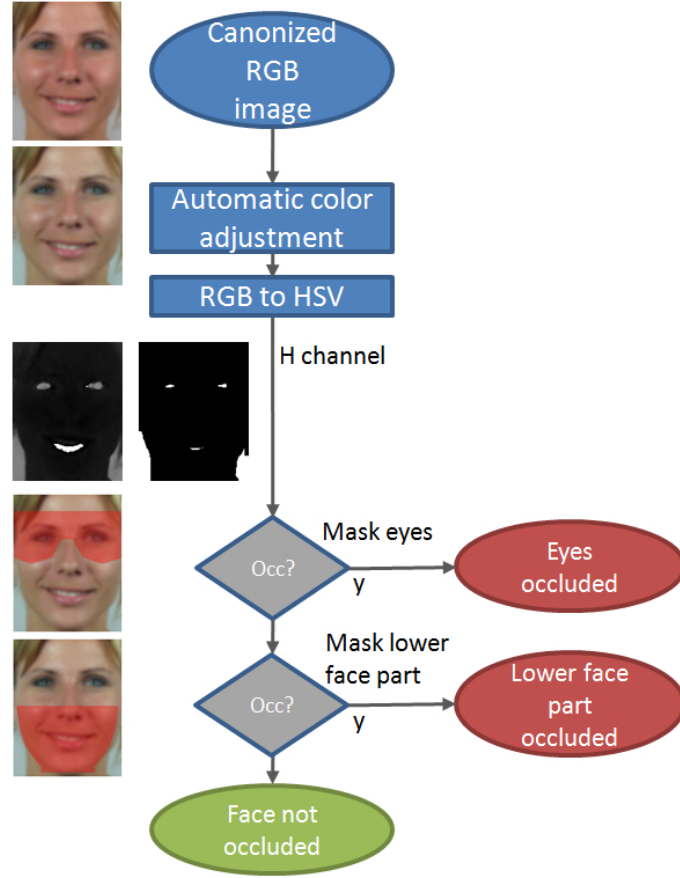


Figure 5.1: Method 1. The tokenized and color adjusted image is transformed to the HSV color space. After binarization of the H-channel of the HSV color space, occlusion masks are used to calculate the level of occlusion on the lower facial part and around the eyes region.

where \bar{Y} is the mean value of the luminance channel and \bar{R} , \bar{G} and \bar{B} correspond to the mean values of the three planes of an RGB-image.

Based on several experiments using different color spaces, e.g., RGB, YUV, LAB, XYZ, YCbCr, we found that the H-channel (representing the hue values of the image) of the HSV color space is best suited for occlusion detection on facial images. The H-channel image is binarized and some morphological post-processing is applied to remove small isolated regions. We define masks for the lower facial part and the eyes to obtain a final value for the level of occlusion. This whole chain is illustrated in Figure 5.1. Figure 5.2 shows some examples of extracting the H-channel of an image and the final occlusion map after binarization. Note that almost always the beards of male individuals are part of the non-occluded facial region in the binarized H-channel image, thus they do not contribute

to an occlusion area.

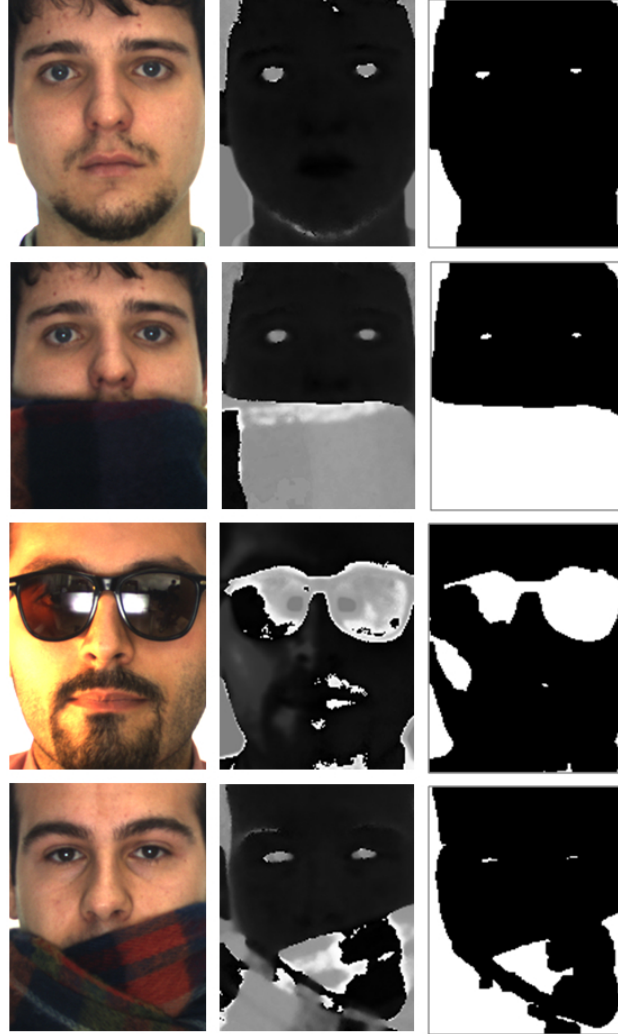


Figure 5.2: Creation of an occlusion map based on the HSV color space. (first column) original images, (second column) H channel of HSV color space and (third column) the corresponding maps gained after thresholding the H channel image.

Experimental results Using Method 1 we conducted experiments on the publicly available AR face database [90]. The AR face database consists of more than 3000 frontal view facial color images of 135 people showing variations in gender, facial expression, illumination conditions and occlusions (sun glasses and scarves). The size of the images is 768×576 pixels. Those individuals wearing a scarf or sun glasses are labeled as occluded. Some representative examples are presented in Figure 5.3. With Method 1 we reached an equal

error rate (EER) of 4.5%. The corresponding ROC curve is shown in Figure 5.4. Note that these results are slightly worse than the results presented in Oh et al. [104], however, their method explicitly trains on the occlusions of the AR face database, which is rather unrealistic, since real occlusions occur in a significantly larger variety, while our method works completely unsupervised. Given the very restricted set of possible occlusions present in the AR face database, we conclude that one needs a database with more variation in order to assess the occlusion detection performance realistically.

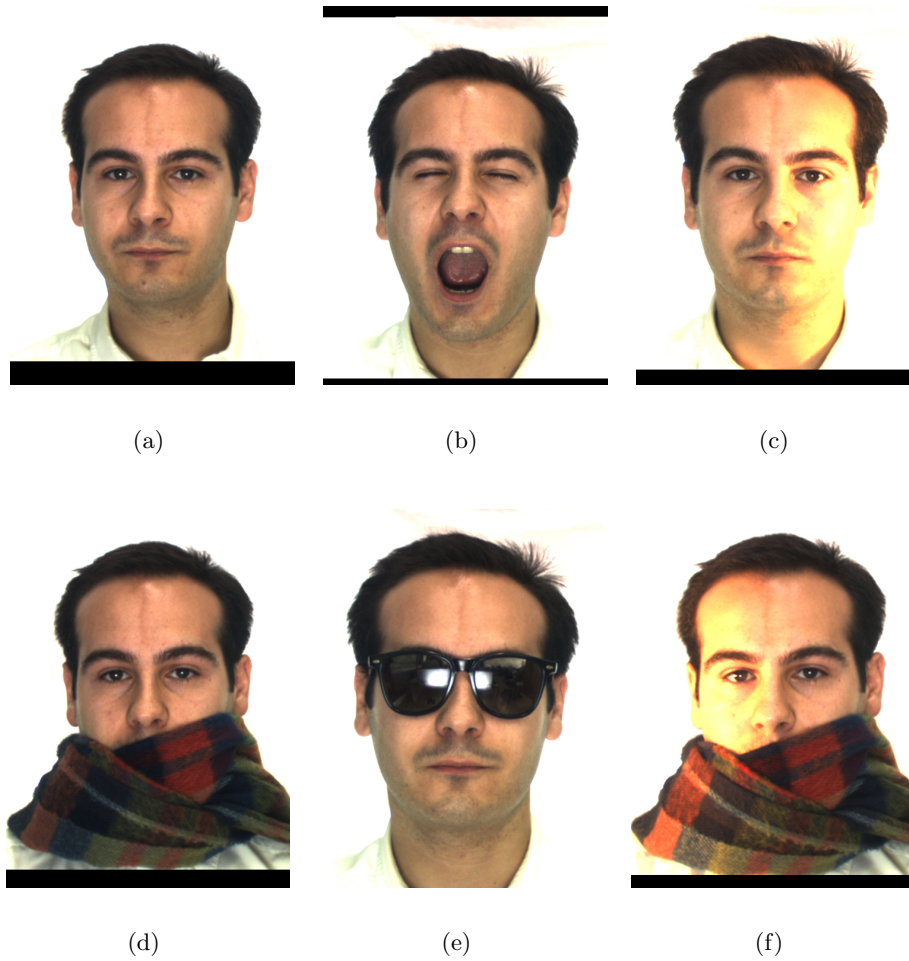


Figure 5.3: Samples from the AR database. The images feature frontal view faces with (a)-(b) different facial expressions, (c) several illumination conditions, (d) occlusion of the lower facial part by a scarf, (e) occlusion of the eyes by sunglasses and (f) combinations of these variations.

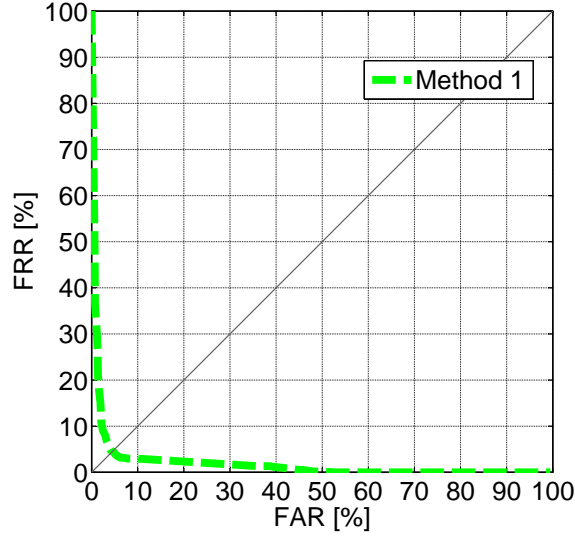


Figure 5.4: ROC curve of Method 1 evaluated on the AR database.

5.1.1.2 Method 2

Method 1 discussed in Section 5.1.1.1 is already well suited to detect occlusions on the publicly available AR database. We created a more challenging database of occluded and non-occluded facial images (see Figure 5.5) where Method 1 does not perform very well. The main reason is that the occlusion detection is exclusively based on color space techniques. In our database we also have skin colored facial occlusions, e.g., hands occluding the face. Examples are depicted in Figure 5.5a and Figure 5.5f where Method 1 would fail.

We start by improving the color detection branch as shown in Figure 5.6. We transform our lowpass filtered tokenized image to the HSV color space and extract the H-channel. After binarization and some morphological operations (as used for Method 1) we obtain the occlusion map. We define an occlusion mask for the forehead and the lower facial part. The eyes region will be considered as the method proceeds. The forehead mask is used to calculate the level of occlusion of the forehead. This is especially important if e.g., somebody wears a cap (Figure 5.5b).

If the approach at that stage claims an occlusion of the lower facial part, we validate this claim by finding similar colors based on a given color mask, Figure 5.7. The color mask C is defined based on the position of the tokenized image and is marked by the red lines in Figure 5.7a. We pick up the H-pixel values $h_j \in C$ and form a unimodal Gaussian model $\mathcal{N}(\mu, \sigma^2)$. It turns out that using only the H-values for constructing the Gaussian



Figure 5.5: Samples from our own database. In addition to the variations exhibited by the AR database, our own database shows some more variations, e.g. (a) occlusions by skin-similar color of the lower facial part, (b) occlusions of the forehead, (c) variation of the color tone of the overall image, (d) extreme lighting conditions, (e) tinted glasses in several colors and (f) several colored occlusions of the lower facial part (also skin-similar color).

model is superior to a multivariate Gaussian model constructed from several color queues. Using the Gaussian model we compute the probability map (Figure 5.7b) in the facial image domain Ω . Therefore, we calculate the probability p_i , $i \in \Omega$, of every pixel ν_i to determine how similar it is to the marked pixels in Figure 5.7a:

$$p_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\nu_i - \mu}{\sigma}\right)^2\right). \quad (5.2)$$

After binarization and some morphological operations we get the final facial map (Figure 5.7c).

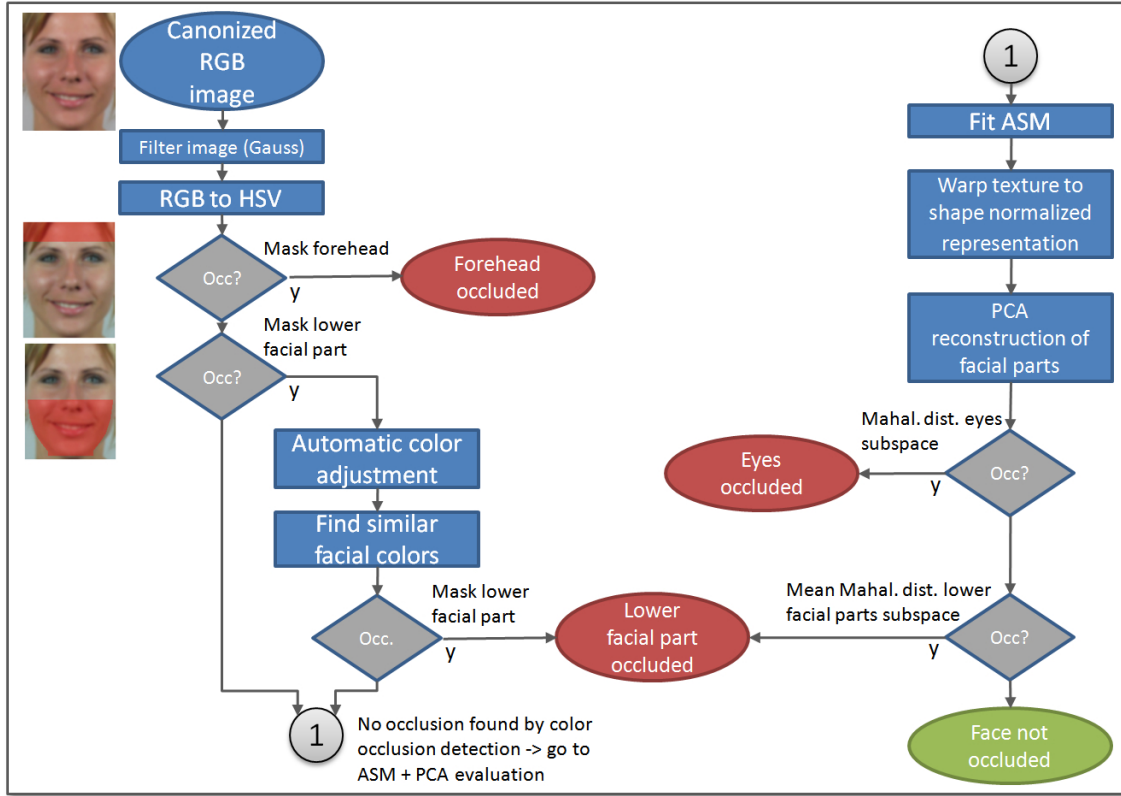


Figure 5.6: Method 2. The left branch of the whole approach is based on color techniques. If there is no occlusion found by this color occlusion detection, the ASM + PCA approach will be activated.

At the end of the color occlusion detection chain, the images showing occlusions with colors similar to skin are still classified as non-occluded. Hence, we create a second occlusion detection step based on fitting an Active Shape Model (ASM) [23], which should also detect non-facial structures. Here we use the recently proposed STASM [93] algorithm which is very robust against varying illumination conditions or partly occluded facial images and showed excellent performance on our data. We also performed experiments with Active Appearance Models [22, 124] but they are by far inferior in terms of fitting accuracy for our task compared to STASM, see Figure 5.8.

The STASM Algorithm This publicly available algorithm extends the original Active Shape Model [23] by a number of techniques like two- instead of one-dimensional landmark

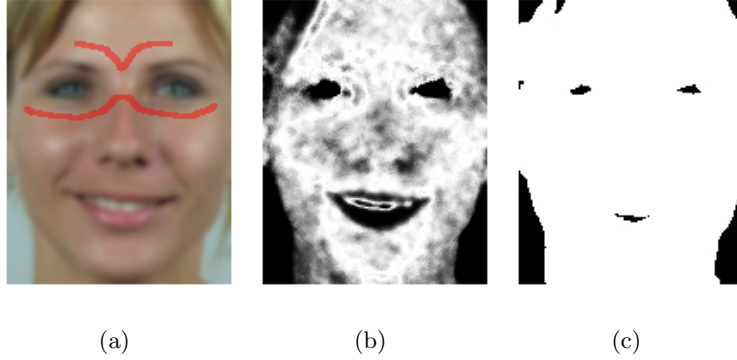


Figure 5.7: Determining similar colors. (a) The H-channel pixel values marked by the red lines are used to construct a Gaussian model. (b) Probability map of similar colors, (c) probability map after binarization and some morphological operations.

profiles, extending the set of training landmarks and trimming the covariance matrix by setting a large number of entries to zero. In the following we will describe this algorithm which is an important part of our system.

The original ASM makes use of a statistical formulation to combine a set of user-specified landmark points in a training set of annotated images into a generative model of the object of interest. This generative model describes the variation of the object shape from a mean object instance. The ASM relies on a specified ordering of the n landmarks $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ in a training image. Given K suitably aligned training images we can generate K vectors \mathbf{x}_k . These vectors

$$\mathbf{x}_k = (x_1, \dots, x_n, y_1, \dots, y_n)_k^T$$

form a distribution in a $2n$ dimensional space, and the aim of the ASM is to generatively model this distribution. Therefore, a Principal Component Analysis (PCA) is applied to the training data which results in the mean and the main axes with their corresponding variances of the cloud of points in the high-dimensional space. An approximation of any training instance \mathbf{x} can be calculated from

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}, \quad (5.3)$$

where $\bar{\mathbf{x}}$ is the mean of the distribution, \mathbf{P} is the matrix formed by the t eigenvectors of the covariance matrix of the points and \mathbf{b} is a t -dimensional vector of weights which resembles a set of parameters of a deformable shape model. Modifying this parameter

creates different shapes restricted by the information from the training data. In Cootes et al. [23] the fitting of the shape model is performed by an iterative algorithm that finds global pose as well as model parameters \mathbf{b} . The fitting procedure is based on the matching of a one-dimensional profile of gray-value and edge information, derived from the training data, to profiles extracted from the current position of the landmarks in the test image. This procedure is very sensitive to the initialization of global pose and model parameters and is prone to get stuck in local minima. The fitting process can be understood as iteratively moving landmark points independently from each other to locations where the profile match is a better one and regularizing the locations of all landmark points by the global PCA shape model.

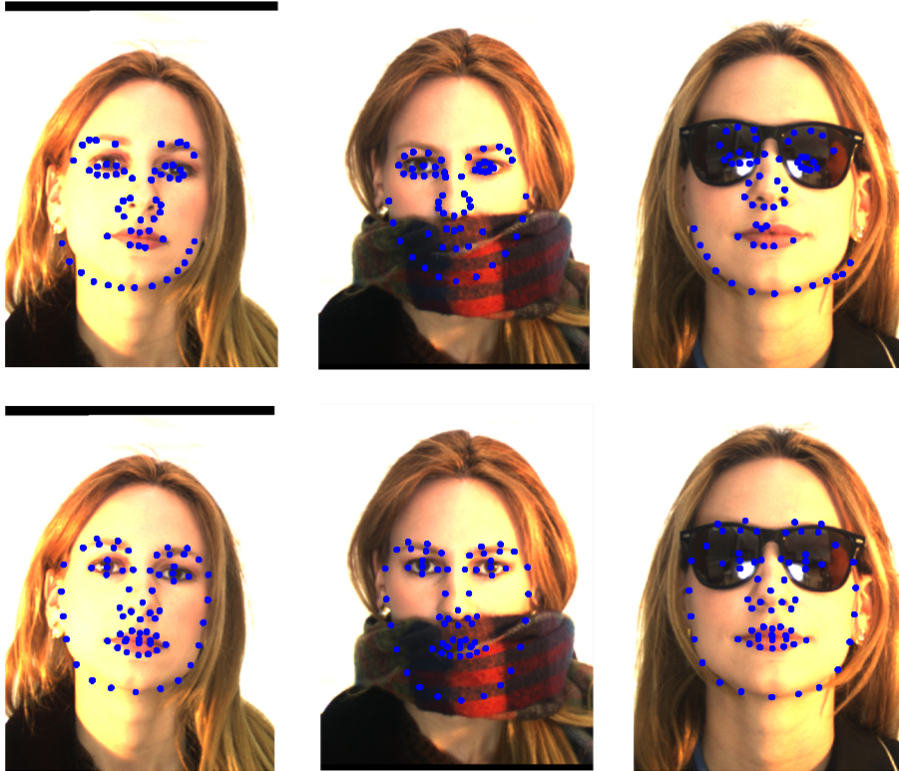


Figure 5.8: Comparison of the fitting quality of model based fitting. (first row) AAM, (second row) STASM.

STASM [93] advances this basic model by a number of important extensions. First, they increase the number of necessary landmark points to add redundancy to the model representation and they perturb the landmarks of the training data set by random noise to increase the number of available training data. Second, instead of one-dimensional profiles they use two-dimensional patches at the landmarks. This increases the matching

performance at the cost of slightly more computational work. Third, during iterative fitting the global shape model is used with an increasing amount of variation. This means that at lower levels in the image pyramid a small variance around the mean shape is allowed and a more restricted set of eigenvalues is chosen for shape regularization. As fitting proceeds and we reach the original level of the image pyramid eigenvectors are added and the maximal variance is increased in order to loosen the regularization constraints imposed by the shape model. Finally, the patch profile covariance matrix used for matching is optimized by setting components resembling distant points to zero and trimming the resulting approximated covariance in order to be positive definite again. The main purpose of this step is to reduce matching time.

Combining STASM with a PCA sub-component model We manually annotated n facial images and aligned the obtained facial shapes by applying Procrustes analysis for shape registration. The mean shape is calculated from these aligned shapes. We warp each annotated image to the mean shape representation and split every warped image into three parts, namely the left and right part of the lower face and the eyes region, see Figure 5.9. We construct a separate color Principal Component Analysis (PCA) subspace $\mathbf{U}_k = [\mathbf{u}_{k_1}, \dots, \mathbf{u}_{k_p}]$, $k \in [1, 2, 3]$ for every part. Usually only p , $p < n$, eigenvectors \mathbf{u} are sufficient.

In the occlusion detection with this ASM + PCA approach we first fit the ASM to the input image. We warp the texture, enclosed in the found landmarks, to the mean shape representation. This texture is split as in the training stage of the PCA. The advantage of the split is the increased robustness to bad illumination conditions compared to the whole face. Every obtained facial part \mathbf{t}_k is projected into the corresponding subspace obtaining the PCA coefficients \mathbf{c}_k which correspond to distances from the mean on the axes spanned by the subspace:

$$\mathbf{c}_k = \mathbf{U}_k^T (\mathbf{t}_k - \overline{\mathbf{t}_k}). \quad (5.4)$$

We measure the Mahalanobis distance d_k in every partial subspace and thus determine which part of the face is occluded:

$$d_k = \sqrt{\mathbf{c}_k^T \boldsymbol{\Sigma}_k^{-1} \mathbf{c}_k}, \quad (5.5)$$

where $\boldsymbol{\Sigma}_k = \text{diag}(\lambda_k)$ is the diagonal covariance matrix consisting of the eigenvalues λ_k obtained from the construction of the PCAs.

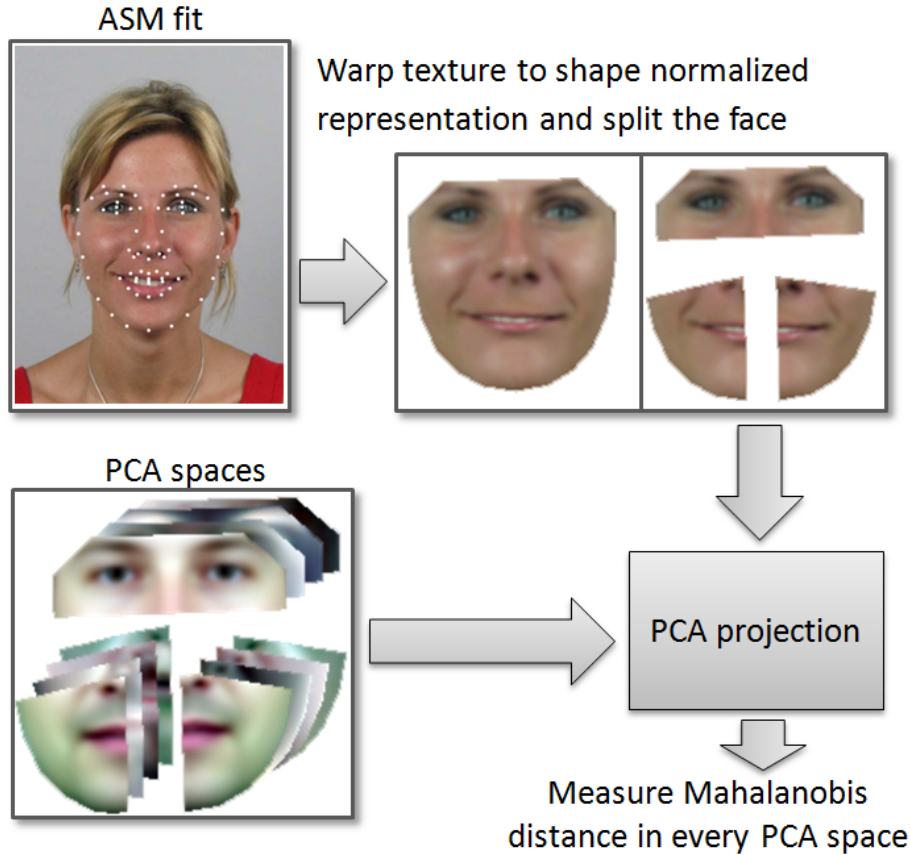


Figure 5.9: ASM + PCA for occlusion detection.

A further nice property of this ASM + PCA approach is that we can measure the fitting quality of the ASM, which is an unresolved question in the literature. If the ASM fit is poor, the warped and projected texture will lead to a large Mahalanobis distance in the subspace, because this texture is not represented in the facial subspace. On the other hand a good ASM fit will result in a good reconstruction of the facial parts. Hence, that combined approach is a good indicator for the fitting quality of the ASM.

Experimental results We performed experiments on our own database which consists of 4930 color facial images and is more challenging compared to the AR database used in our first experiments. In addition to the variations of the AR database, our database exhibits further illumination conditions and more types of occlusions, see Figure 5.5. The size of the images is 480×640 pixels.

For our ASM + PCA approach, we manually annotated 427 facial images taken from the Caltech face database [19] and our own collection (disjoint from our test database).

Taking also the mirrored versions of those images doubles the amount of data. For the PCA model we keep 98% of the eigenvalue energy spectrum for each of the three subspaces.

Method 2 gives a significant increase in performance compared to Method 1 on our own database. The EER is decreased from 30.9% to 6.6%. The corresponding ROC curves are depicted in Figure 5.10. In Figure 5.11 some typical failure cases of our approach are shown. The algorithm is very fast, mostly depending on the runtime of the STASM. The average runtime* to analyze a facial image is 0.3s.

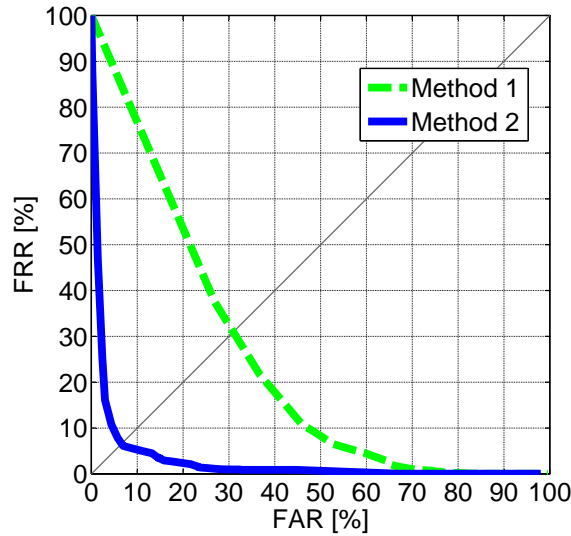


Figure 5.10: ROC curve of Method 1 and Method 2 evaluated on our own dataset.

5.1.2 Summary

Occlusion detection is an important part of the ICAO specification for assessing suitability of facial images for machine readable travel documents. We presented two approaches which detect occlusions on facial portrait images. The first approach is straightforward and is based on color techniques using the H-channel of the HSV color space. It turns out that this first method is already well suited to sufficiently detect occlusions on the publicly available AR database. We created a more challenging database where the first method showed significant shortcomings. Hence, we improved our first method with a combination of an Active Shape Model (ASM) and a component based PCA subspace reconstruction. This algorithm proved to be very successful on our more difficult database. Furthermore, we can use the combination of ASM and PCA reconstruction as a measure of ASM fitting

*The runtime is measured using an Intel Core 2 Duo processor running at 2.4GHz.



Figure 5.11: Typical failure cases resulting from (a)-(c) skin-similar occlusion of the forehead, (d) larger deviations from frontal pose, (e) extreme facial hair and (f) slight specularities exhibited on the glasses.

quality.

5.2 Active Appearance Model Fitting Under Occlusion

In the previous section we showed how to detect an occlusion. In the case of an occluded input image, we can also improve the fitting quality of an Active Appearance Model (AAM). Generative model-based approaches for feature localization have received a lot of attention over the last decade. Their key advantage is to use a priori knowledge from a training stage for restricting the model while searching for a model instance in an image. Two specific instances of model-based approaches, the Active Appearance Model (AAM) [22] and the closely related 3D Morphable Model (3DMM) [15], have proven to show excellent results in locating image features in applications such as face detection and tracking [91], face and facial expression recognition [16], or medical image segmentation [7, 94].

Despite its large success, the AAM model has one main limitation. It is not robust against occlusions. Thus, if important features are missing the AAM fitting algorithm tends to get stuck in local minima. This especially credits for human faces since the large variability in the image data such as certain kinds of glasses, makeup, or beards cannot totally be captured in the training stage. Similar difficulties also arise in other areas of model-based approaches (e.g., in the medical domain [7]).

In the recent years some research was dedicated to generative model-based approaches in the presence of occlusions by investigating robust fitting strategies. In the original AAM approach [22] fitting is treated as a least squares optimization problem, which is, of course, very sensitive to outliers due to its quadratic error measure (L_2 norm). To overcome this problem, the work of [35] extended the standard fitting method (a) by learning the usual gray-value differences encountered during training and (b) by ignoring gray-value differences exceeding a threshold derived from these values during fitting. But the main drawback of this method is that the required threshold depends on the training conditions, which makes it improper for real-life situations. In contrast, in [32] a RANSAC procedure is used for the initialization of the AAM fitting in order to get rid of occlusions due to differing poses. However, since the AAM fitting remains unchanged this approach has still problems with appearance outliers.

Another direction of research was dedicated to replacing the least-squares error measure by a robust error measure in the fitting stage [47]. Later this approach was further refined by comparing several robust error measures [131]. The same strategy is also used in [116] and was adapted to a statistical framework in [157]. But the latter approach is limited in several ways: (a) a scale parameter is required, which is hard to determine in general, (b) the framework around the inverse compositional algorithm is specifically tailored to

tracking, and (c) the face models are built from the tracked person, which limits its applicability for general applications.

In the context of medical image analysis a robust AAM fitting approach was presented in [7]. In their method, which is based on the standard AAM fitting algorithm, gross disturbances (i.e., outliers) in the input image are avoided by ignoring misleading coefficient updates in the fitting stage. For that purpose, inlier and outlier coefficients are identified by a Mean Shift based analysis of the residual's modes. Then, an optimal subset of modes is selected and only those pixels covered by the selected mode combination are used for actual residual calculation. The Robust AAM Matching (RAAM) approach shows excellent results on a number of medical data sets. However, the mode selection is computationally very complex. Thus, this method is impractical for real-time or near real-time applications.

To overcome these drawbacks we introduce a new efficient robust AAM fitting scheme. In contrast to existing methods the robustness (against occluded features) is not directly included in the fitting step but is detached. In fact, we propose to run a robust pre-processing step first to generate undisturbed input data and then to apply a standard AAM fitting. Since the robust step, which is usually computationally intensive, has to be performed only once (and not iteratively in the fitting process), the computational costs can be reduced.

In particular, the main idea is to robustly replace the missing feature information from a reliable model. Thus, our work is somehow motivated by [101] and [33], where beards and eye-glasses, which are typical problems when applying an AAM approach, are removed. In [33] a PCA model was built from facial images that do not contain any eye-glasses. Then, in the removal step the original input images are reconstructed and the regions with the largest reconstruction errors are identified. These pixels are iteratively replaced by the reconstruction. But this approach can only be applied if the absolute number of missing pixels is quite small. In contrast, in [101] two models are computed in parallel, one for bearded faces and one for non-bearded faces. Then, in the removal step for a bearded face the detected beard region is reconstructed from the non-bearded space.

Since both methods are restricted to special types of occlusion or limited by a pre-defined error level, they cannot be applied for general tasks. Thus, in our approach we apply a robust PCA model (e.g., [13, 75, 114]) to cope with occlusions in the original input data. For that purpose, in the learning stage a reliable model is estimated from undisturbed data (i.e., without any occlusions), which is then applied to robustly recon-

struct unreliable values from the disturbed data. However, a drawback of these methods is their computational complexity (i.e., iterative algorithms, multiple hypothesis, etc.), which hinders practical applicability. Thus, we developed a more efficient robust PCA method (FR-PCA, see Section 5.3) that overcomes this limitation.

Even though the proposed robust AAM fitting is quite general, our main interest is to apply it to facial images. Thus, this application is evaluated in the experiments in detail. However, we also note that it is necessary that the image patch, where the robust PCA is applied has to be roughly aligned with the feature under consideration. In the case of our face localization this can be ensured by using a rough face and facial component detection algorithm inspired by the Viola-Jones algorithm [142]. Moreover, the applied PCA model can handle a wide variability in facial images.

In Section 5.2.1 we introduce our robust AAM fitting algorithm that is based on our novel FR-PCA (Section 5.3) scheme. To demonstrate its benefits, we present experimental results on facial images. Finally, we discuss our findings and summarize our work in Section 5.2.3.

5.2.1 Robust AAM Fitting

Since the parameter updates for the fitting process are estimated from the texture’s residual, the standard AAM is not robust against occlusions. To overcome this limitation, we propose to use our FR-PCA, introduced in Section 5.3, as a pre-processing step to remove disturbances in the input image and to perform the AAM fitting on the thus obtained reconstruction, see Figure 5.12. Occlusions cannot only be of artificial spatially coherent nature, which were taken for the quantitative evaluation of the FR-PCA (Section 5.3), but also in case of facial images, beards or glasses. Those disturbances of facial images influence the quality of the fitting process of AAMs. Thus, for the pre-processing step we trained the FR-PCA using facial images which do not exhibit any disturbances, i.e., no beards and no glasses.

Figure 5.13, which was taken from the Caltech Faces data set [19], demonstrates the whole processing chain for robust AAM fitting under occlusion. Figure 5.13(b) shows the initialization of the AAM on the occluded input image. The rough initialization of the AAM is done using a Viola-Jones face detection approach [142], several AdaBoost-based classifiers for locating eyes and mouth, and a face candidate validation scheme to robustly locate the rough face position as described in Section 2.1.

Figure 5.13(c) demonstrates the converged fit of the AAM on the occluded image

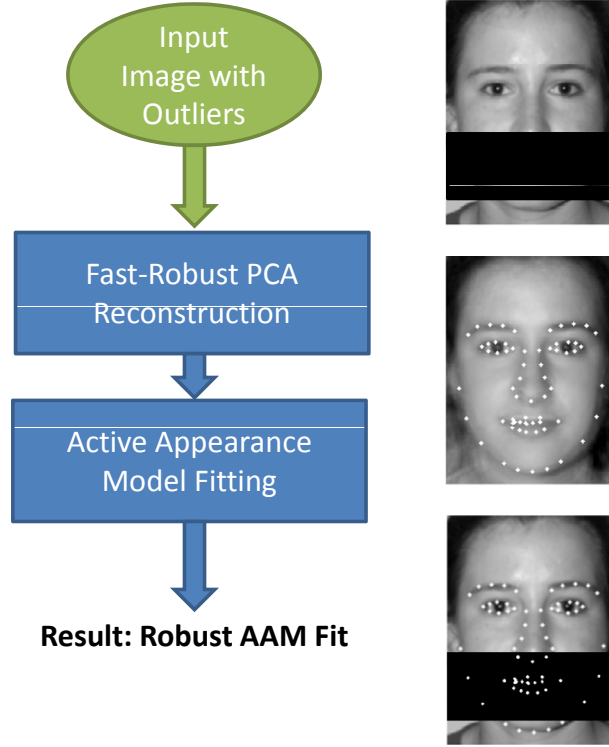


Figure 5.12: Robust AAM fitting chain

which failed totally. In contrast, using the FR-PCA as a pre-processing step results in the converged fit exhibited in Figure 5.13(d). In Figure 5.13(e), the shape from the fitting process on the reconstructed image is overlaid on the original input image. It can be clearly seen that the AAM cannot handle occlusions directly whereas the fit on the reconstructed image is well defined.

5.2.2 Experimental Results

We trained a hierarchical AAM for facial images on three resolution levels (60x80, 120x160, 240x320). Our training set consists of 427 manually annotated face images taken from the Caltech face database [19] and our own collection. Taking also the mirrored versions of those images doubles the amount of training data. For model building we keep 90% of the eigenvalue energy spectrum for the lower two levels and 95% for the highest level to represent our compact model.

As described in Section 5.2.1, we use the FR-PCA as a pre-processing step and perform the AAM fitting on the reconstructed images. Hence, we trained the FR-PCA (Section 5.3.2.1) using facial images which do not exhibit any disturbances, i.e., no beards

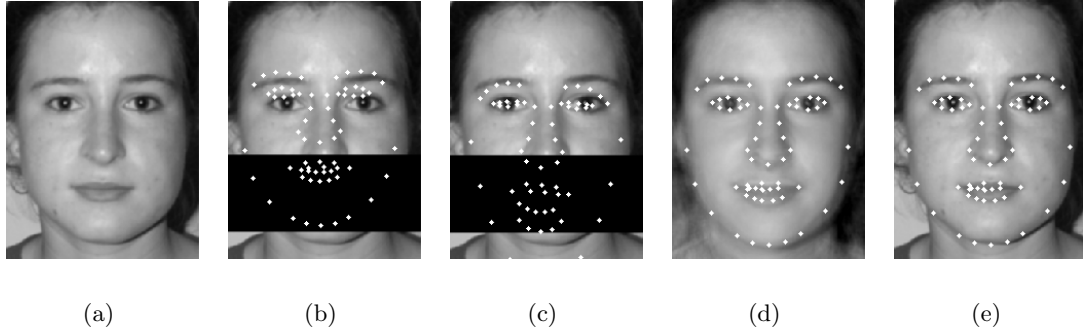


Figure 5.13: Handling of occlusions for AAM fitting. (a) Test image. (b) Initialization of the AAM on the occluded image. (c) Direct AAM fit on occluded image. (d) AAM fit on reconstructed image. (e) Shape from (d) overlaid on the test image. Image taken from Caltech Faces data set [19].

Table 5.1: Point-to-Point error. Comparing the direct fit of the AAM on the test image to the AAM fit utilizing the FR-PCA pre-processing (point errors are measured on 240x320 facial images).

Occlusion	Point-Point Error									
	0%		10%		20%		30%		40%	
	mean	std	mean	std	mean	std	mean	std	mean	std
AAM	4.05	5.77	12.06	11.25	15.19	12.78	18.76	14.89	18.86	13.94
AAM + FR-PCA	5.47	4.97	5.93	5.41	6.06	5.27	9.31	8.75	11.33	9.25

and no glasses. The variance retained for the whole subspace and for the sub-subspaces is 95%.

A 5-fold cross validation is performed using the manually annotated images, resulting in 80% training- and 20% test data per iteration. For each level of occlusion, 210 AAM fits are executed. Table 5.1 shows the point-to-point error (Euclidean distance of converged points to the annotated points) comparing the direct AAM fit on the occluded image to the AAM fit utilizing the FR-PCA pre-processing. Starting from 0% occlusion, the error for the AAM + FR-PCA is slightly larger than the direct fit, because of the unavoidable reconstruction-blur resulting from the FR-PCA reconstruction. When increasing the size of the occlusion, the big advantage of the FR-PCA pre-processing can be seen.

Up to now, to have a steerable environment, we used artificial spatially coherent occlusions. To show the advantage of FR-PCA pre-processing also on natural occlusions such as tinted glasses, occlusions caused by wearing a scarf or by disturbances like beards, Figure 5.14 depicts some AAM fits on images taken from the AR face database [90]. In addition, Figure 5.15 shows an illustrative result on our own database. The FR-PCA

pre-processing step takes around 0.69s per image (150x200) measured in MATLAB using an Intel Xeon processor running at 3GHz.

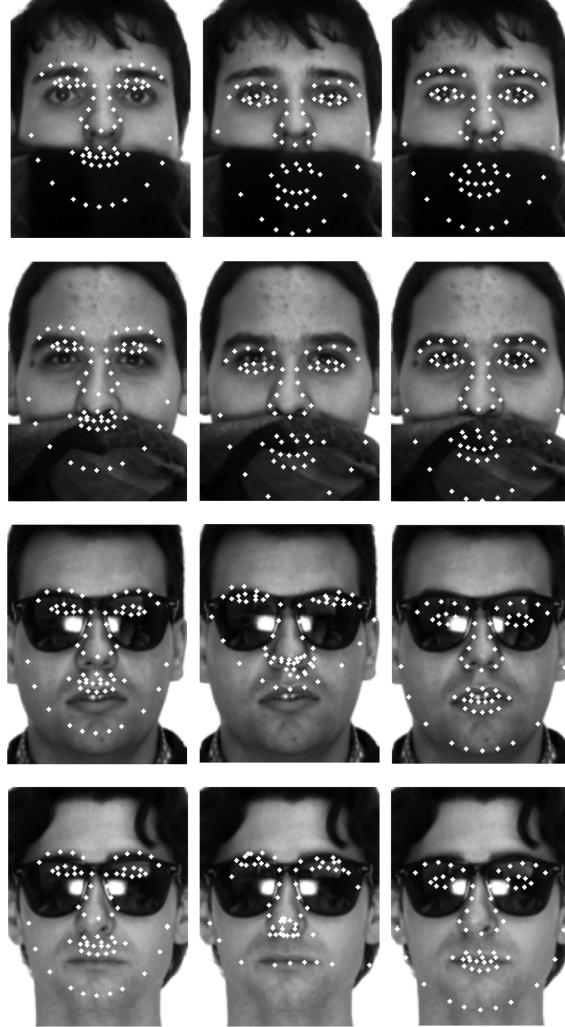


Figure 5.14: Examples of AAM fits on natural occlusions like tinted glasses or wearing a scarf. (First column) Test images with AAM initialization. (Second column) Direct AAM fit on the test images. (Third column) AAM fit utilizing the FR-PCA pre-processing. Images are taken from the AR face database [90].

5.2.3 Summary

We presented a robust method for AAM fitting. In contrast to existing approaches the robustness is not included in the fitting step but is detached in a pre-processing step. The main idea is to robustly reconstruct unreliable data points (i.e., occlusions) in the pre-

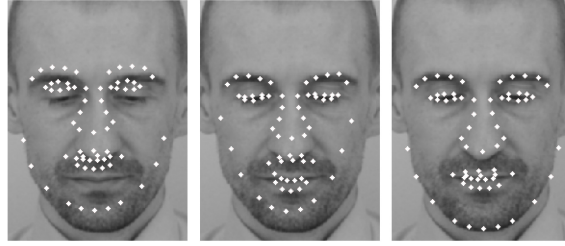


Figure 5.15: Examples of AAM fits on natural occlusions like beards.

processing step and to use the thus obtained undisturbed images as input for a standard AAM fitting. To speed up this robust pre-processing step, we developed a novel fast robust PCA method, see Section 5.3. The whole process chain (robust pre-processing and AAM fitting) was demonstrated in the field of face normalization in the presence of artificial and natural occlusion noise. The results show that our robust approach can handle such situations considerably better than a non-robust approach. Moreover, due to the very efficient robust pre-processing, the proposed robust AAM fitting method is applicable in practice for real-time applications.

5.3 Fast-Robust PCA

Principal Component Analysis (PCA) is a powerful and widely used tool in Computer Vision and is applied, e.g., for dimensionality reduction. But as a drawback, it is not robust to outliers. Hence, if the input data is corrupted, an arbitrarily wrong representation is obtained. To overcome this problem, various methods have been proposed to robustly estimate the PCA coefficients, but these methods are computationally too expensive for practical applications. Thus, in this section we propose a novel fast and robust PCA (FR-PCA), which drastically reduces the computational effort. Moreover, more accurate representations are obtained. In particular, we propose a two-stage outlier detection procedure, where in the first stage outliers are detected by analyzing a large number of smaller subspaces. In the second stage, remaining outliers are detected by a robust least-square fitting. To show these benefits, in the experiments we evaluate the FR-PCA method for the task of robust image reconstruction on the publicly available ALOI database. The results clearly show that our approach outperforms existing methods in terms of accuracy and speed when processing corrupted data. We apply our FR-PCA algorithm to reconstruct missing feature information and to use the obtained image as input for the standard AAM fitting process, see Section 5.2.

5.3.1 Preface

Principal Component Analysis (PCA) [61] also known as Karhunen-Loève transformation (KLT) is a well known and widely used technique in statistics. The main idea is to reduce the dimensionality of data while retaining as much information as possible. This is assured by a projection that maximizes the variance but minimizes the mean squared reconstruction error at the same time. Murase and Nayar [97] showed that high dimensional image data can be projected onto a subspace such that the data lies on a lower dimensional manifold. Thus, starting from face recognition (e.g., [64, 137]) PCA has become quite popular in computer vision*, where the main application of PCA is dimensionality reduction. For instance, a number of powerful model-based segmentation algorithms such as Active Shape Models [23] or Active Appearance Models [22] incorporate PCA as a fundamental building block.

In general, when analyzing real-world image data, one is confronted with unreliable data, which leads to the need for robust methods (e.g., [48, 55]). Due to its least squares formulation, PCA is highly sensitive to outliers. Thus, several methods for robustly learning PCA subspaces (e.g., [118, 123, 134, 136, 152]) as well as for robustly estimating the PCA coefficients (e.g., [13, 36, 75, 114]) have been proposed. We are focusing on the latter case. Thus, in the learning stage a reliable model is estimated from undisturbed data, which is then applied to robustly reconstruct unreliable values from the unseen corrupted data.

To robustly estimate the PCA coefficients Black and Jepson [13] applied an M-estimator technique. In particular, they replaced the quadratic error norm with a robust one. Similarly, Rao [114] introduced a new robust objective function based on the MDL principle. But as a disadvantage, an iterative scheme (i.e., EM algorithm) has to be applied to estimate the coefficients. In contrast, Leonardis and Bischof [75] proposed an approach that is based on sub-sampling. In this way, outlying values are discarded iteratively and the coefficients are estimated from inliers only. Similarly, Edwards and Murase introduced adaptive masks to eliminate corrupted values when computing the sum-squared errors.

A drawback of these methods is their computational complexity (i.e., iterative algorithms, multiple hypotheses, etc.), which limits their practical applicability. Thus, we develop a more efficient robust PCA method that overcomes this limitation. In particular,

*For instance, at CVPR 2007 approximative 30% of all papers used PCA at some point (e.g., [74, 129, 144]).

we propose a two-stage outlier detection procedure. In the first stage, we estimate a large number of smaller subspaces sub-sampled from the whole dataset and discard those values that are not consistent with the subspace models. In the second stage, the data vector is robustly reconstructed from the thus obtained subset. Since the subspaces estimated in the first step are quite small and only a few iterations of the computationally more complex second step are required (i.e., most outliers are already discarded by the first step), the whole method is computationally very efficient. This is confirmed by the experiments, where we show that the proposed method outperforms existing methods in terms of speed and accuracy.

This section is structured as follows. In Section 5.3.2, we introduce and discuss the novel fast-robust PCA (FR-PCA) approach. Experimental results for the publicly available ALOI database are given in Section 5.3.3. Finally, we discuss our findings and summarize our work in Section 5.3.4.

5.3.2 Derivation of the Algorithm

Given a set of n high-dimensional data points $\mathbf{x}_j \in \mathbb{R}^m$ organized in a matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, then the PCA basis vectors $\mathbf{u}_1, \dots, \mathbf{u}_{n-1}$ correspond to the eigenvectors of the sample covariance matrix

$$\mathbf{C} = \frac{1}{n-1} \hat{\mathbf{X}} \hat{\mathbf{X}}^T, \quad (5.6)$$

where $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n]$ is the mean normalized data with $\hat{\mathbf{x}}_j = \mathbf{x}_j - \bar{\mathbf{x}}$. The sample mean $\bar{\mathbf{x}}$ is calculated by

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j. \quad (5.7)$$

Given the PCA subspace $\mathbf{U}_p = [\mathbf{u}_1, \dots, \mathbf{u}_p]$ (usually only p , $p < n$, eigenvectors are sufficient), an unknown sample $\mathbf{x} \in \mathbb{R}^m$ can be reconstructed by

$$\tilde{\mathbf{x}} = \mathbf{U}_p \mathbf{a} + \bar{\mathbf{x}} = \sum_{j=1}^p a_j \mathbf{u}_j + \bar{\mathbf{x}}, \quad (5.8)$$

where $\tilde{\mathbf{x}}$ denotes the reconstruction and $\mathbf{a} = [a_1, \dots, a_p]$ are the PCA coefficients obtained by projecting \mathbf{x} onto the subspace \mathbf{U}_p .

If the sample \mathbf{x} contains outliers, (5.8) does not yield a reliable reconstruction; a robust method is required (e.g., [13, 36, 75, 114]). But since these methods are computationally

very expensive (i.e., they are based on iterative algorithms) or can handle only a small amount of noise, they are often not applicable in practice. Thus, in the following we propose a new fast robust PCA approach (FR-PCA), which overcomes these problems.

5.3.2.1 FR-PCA Training

The training procedure, which is sub-divided into two major parts, is illustrated in Figure 5.16. First, a standard PCA subspace \mathbf{U} is generated using the full available training data. Second, N sub-samplings \mathbf{s}_n are established from randomly selected values from each data point (illustrated by the red points and the green crosses). For each sub-sampling \mathbf{s}_n , a smaller subspace (sub-subspace) \mathbf{U}^n is estimated, in addition to the full subspace.

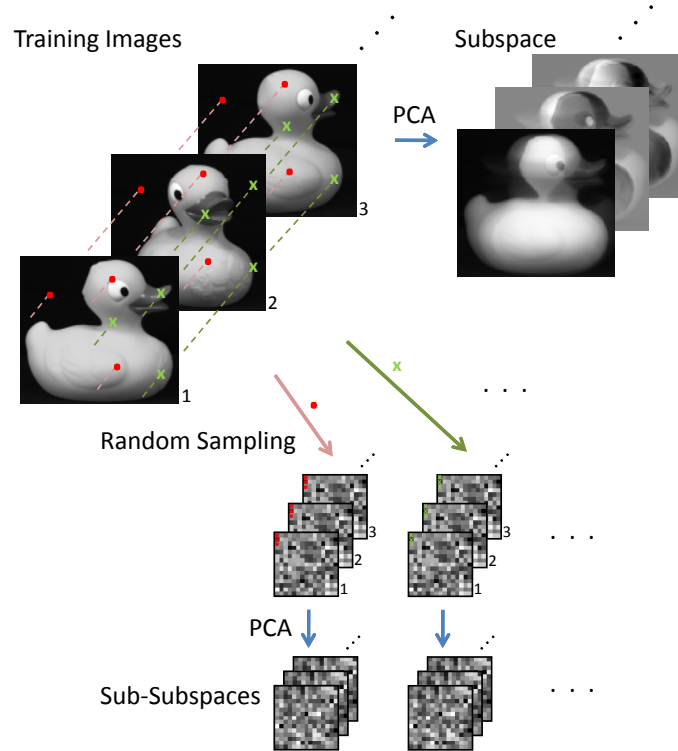


Figure 5.16: FR-PCA training: A global PCA subspace and a large number of smaller PCA sub-subspaces are estimated in parallel. Sub-subspaces are derived by randomly sub-sampling the input data.

Since occlusions are mainly considered to be spatially coherent the sub-sampling is done in a smart way. In addition to the random sampling over the whole image region, further random samplings are also restricted to image slices (vertical, horizontal and quadrant). This is illustrated in Figure 5.17.

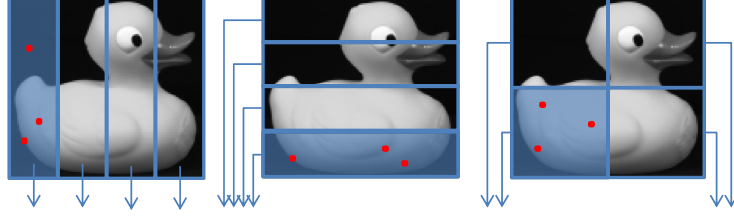


Figure 5.17: Random sampling restricted to image slices (vertical, horizontal and quadrant).

5.3.2.2 FR-PCA Reconstruction

Given a new unseen test sample \mathbf{x} , the robust reconstruction $\tilde{\mathbf{x}}$ is estimated in two stages. In the first stage (*gross outlier detection*), the outliers are detected based on the reconstruction errors of the sub-subspaces. In the second stage (*refinement*), using the thus estimated inliers, a robust reconstruction $\tilde{\mathbf{x}}$ of the whole sample is generated, see Figure 5.18.

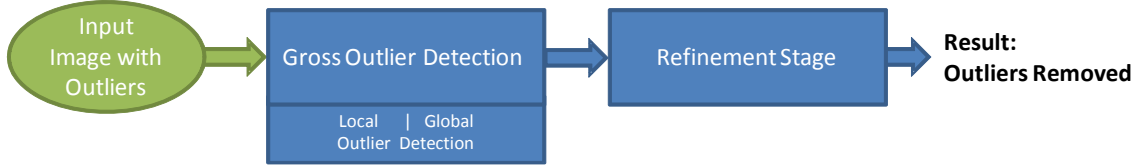


Figure 5.18: Reconstruction pipeline

In the *gross outlier detection*, first, N sub-samplings \mathbf{s}_n are generated according to the corresponding sub-subspaces \mathbf{U}^n , which were estimated as described in Section 5.3.2.1. In addition, we define the set of “inliers” \mathbf{r} as the union of all selected pixels: $\mathbf{r} = \mathbf{s}_1 \cup \dots \cup \mathbf{s}_N$, which is illustrated in Figure 5.20(a) (green points). Next, for each sub-sampling \mathbf{s}_n a reconstruction $\tilde{\mathbf{s}}_n$ is estimated by (5.8), which allows to estimate the error-maps

$$\mathbf{e}_n = |\mathbf{s}_n - \tilde{\mathbf{s}}_n|, \quad (5.9)$$

the mean reconstruction error \bar{e} over all sub-samplings, and the mean reconstruction errors \bar{e}_n for each of the N sub-samplings.

Based on these errors, we can detect the outliers by local and global thresholding. The local thresholds (one for each sub-sampling) are defined by $\theta_n = \bar{e}_n w_n$, where w_n is a weighting parameter and the global threshold θ is set to the mean error \bar{e} . Then, all

points $s_{n,(i,j)}$ for which

$$e_{n,(i,j)} > \theta_n \quad \text{or} \quad e_{n,(i,j)} > \theta \quad (5.10)$$

are discarded from the sub-samplings \mathbf{s}_n obtaining $\hat{\mathbf{s}}_n$. Finally, we re-define the set of “inliers” by

$$\mathbf{r} = \hat{\mathbf{s}}_1 \cup \dots \cup \hat{\mathbf{s}}_q, \quad (5.11)$$

where $\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_q$ indicate the first q sub-samplings (sorted by \bar{e}_n) such that $|\mathbf{r}| \leq k$; k is the pre-defined maximum number of points. The thus obtained “inliers” are shown in Figure 5.20(b).

The *gross outlier detection* procedure allows to remove most outliers. Thus, the obtained set \mathbf{r} contains almost only inliers. To further improve the final result in the *refinement* step, the final robust reconstruction is estimated similar to [75]. Starting from the point set $\mathbf{r} = [r_1, \dots, r_k]$, $k > p$, obtained from the *gross outlier detection*, repeatedly reconstructions $\tilde{\mathbf{x}}$ are computed by solving an over-determined system of equations (Figure 5.19) minimizing the least squares reconstruction error

$$E(\mathbf{r}) = \sum_{i=1}^k \left(x_{r_i} - \sum_{j=1}^p a_j \mathbf{u}_{j,r_i} \right)^2. \quad (5.12)$$

Thus, in each iteration those points with the largest reconstruction errors can be discarded from \mathbf{r} (selected by a reduction factor α). These steps are iterated until a pre-defined number of remaining points is reached. Finally, an outlier-free subset is obtained, which is illustrated in Figure 5.20(c).

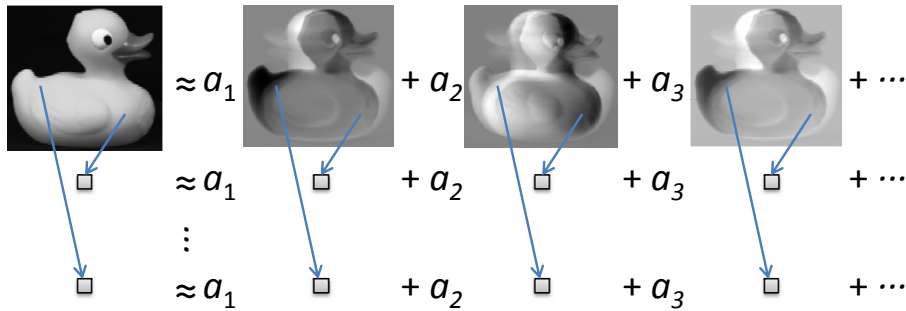


Figure 5.19: Refinement step. Solve an over-determined system of equations.

A robust reconstruction result obtained by the proposed approach compared to a non-robust method is shown in Figure 5.21. One can clearly see that the robust method considerably outperforms the standard PCA. Note, the blur visible in the reconstruc-

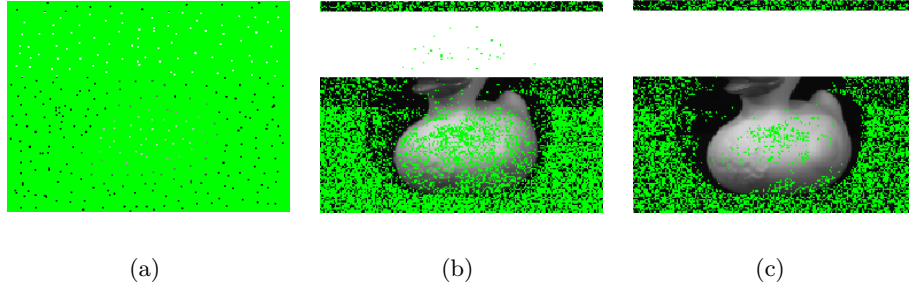


Figure 5.20: Data point selection process: (a) data points sampled by all sub-subspaces, (b) occluded image showing the remaining data points after applying the sub-subspace procedure, and (c) resulting data points after the iterative refinement process for the calculation of the PCA coefficients. This figure is best viewed in color.

tion of the FR-PCA is the consequence of taking into account only a limited number of eigenvectors.

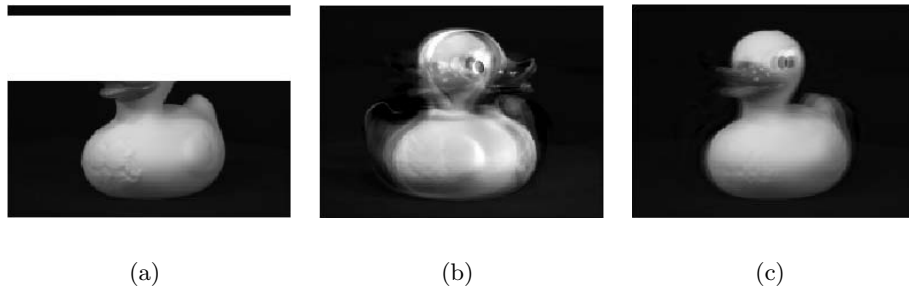


Figure 5.21: Demonstration of the insensitivity of the robust PCA to noise (i.e., occlusions): (a) occluded image, (b) reconstruction using standard PCA, and (c) reconstruction using the FR-PCA.

In general, the robust estimation of the coefficients is computationally very efficient. In the *gross outlier detection* procedure, only simple matrix operations have to be performed, which are very fast; even if hundreds of sub-subspace reconstructions have to be computed. The computationally more expensive part is the *refinement* step, where repeatedly an overdetermined linear system of equations has to be solved. Since only very few refinement iterations have to be performed due to the preceding *gross outlier detection*, the total runtime is kept low.

5.3.3 Experimental Results

To show the benefits of the proposed fast robust PCA method (FR-PCA), we compare it to the standard PCA (PCA) and the robust PCA approach presented in [75] (R-PCA). We choose the latter one, since it yields superior results among the presented methods in the literature and our refinement process is similar to theirs.

In particular, the experiments are evaluated for the task of robust image reconstruction on the “Amsterdam Library of Object Images (ALOI)” database [46]. The ALOI database consists of 1000 different objects. Over hundred images of each object are recorded under different viewing angles, illumination angles and illumination colors, yielding a total of 110,250 images. For our experiments we arbitrarily choose 30 categories (009, 018, 024, 032, 043, 074, 090, 093, 125, 127, 135, 138, 151, 156, 171, 174, 181, 200, 299, 306, 323, 354, 368, 376, 409, 442, 602, 809, 911, 926), where an illustrative subset of objects is shown in Figure 5.22.



Figure 5.22: Illustrative examples of ALOI database objects [46] used in the experiments.

In our experimental setup, each object is represented in a separate subspace and a set of 1000 sub-subspaces, where each sub-subspace contains 1% of data points of the whole image. The variance retained for the sub-subspaces is 95% and 98% for the whole subspace, which is also used for the standard PCA and the R-PCA. Unless otherwise noted, all experiments are performed with the parameter settings given in Table 5.2.

A 5-fold cross-validation is performed for each object category, resulting in 80% training- and 20% test data, corresponding to 21 test images per iteration. The experiments are accomplished for several levels of spatially coherent occlusions and

Table 5.2: Parameters for the FR-PCA (a) and the R-PCA (b) used for the experiments.

(a)		(b)	
FR-PCA		R-PCA	
Number of initial points k	$130p$	Number of initial hypotheses H	30
Reduction factor α	0.9	Number of initial points k	$48p$
		Reduction factor α	0.85
		K_2	0.01
		Compatibility threshold	100

several levels of salt & pepper noise. Quantitative results for the root-mean-squared (RMS) reconstruction-error per pixel for several levels of occlusions are given in Table 5.3. In addition, in Figure 5.23 we show box-plots of the RMS reconstruction-error per pixel for different levels of occlusions. Analogously, the RMS reconstruction-error per pixel for several levels of salt & pepper noise is presented in Table 5.4 and the corresponding box-plots are shown in Figure 5.24.

Table 5.3: Comparison of the reconstruction errors of the standard PCA, the R-PCA and the FR-PCA for several levels of occlusion showing RMS reconstruction-error per pixel given by (a) mean and standard deviation and (b) median, lower- and upper quartile.

(a)												
Occlusion	Error per Pixel											
	0%		10%		20%		30%		50%		70%	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
PCA	10.06	6.20	21.82	8.18	35.01	12.29	48.18	15.71	71.31	18.57	92.48	18.73
R-PCA	11.47	7.29	11.52	7.31	12.43	9.24	22.32	21.63	59.20	32.51	94.75	43.13
FR-PCA	10.93	6.61	11.66	6.92	11.71	6.95	11.83	7.21	26.03	23.05	83.80	79.86

(b)																		
Occlusion	Error per Pixel																	
	0%			10%			20%			30%			50%			70%		
	median	Q ₂₅	Q ₇₅	median	Q ₂₅	Q ₇₅	median	Q ₂₅	Q ₇₅	median	Q ₂₅	Q ₇₅	median	Q ₂₅	Q ₇₅	median	Q ₂₅	Q ₇₅
PCA	9.37	5.34	14.32	21.82	16.85	26.49	34.93	27.86	42.47	48.14	38.10	58.77	71.44	58.43	84.45	92.28	78.68	106.59
R-PCA	10.57	6.03	16.20	10.59	6.20	16.37	11.13	6.43	16.67	14.59	8.16	27.44	64.01	31.13	82.96	91.00	73.28	109.31
FR-PCA	10.23	6.13	15.34	11.11	6.71	16.37	11.15	6.74	16.29	11.08	6.69	16.64	17.37	9.35	36.82	76.70	57.02	96.14

From Table 5.3 and Figure 5.23 it can be seen – starting from an occlusion level of 0% – that all subspace methods exhibit nearly the same RMS reconstruction-error. Increasing the portion of occlusion, the standard PCA shows large errors whereas the robust methods are still comparable to the non-disturbed (best feasible) case, where our novel FR-PCA presents the best performance. In contrast, as can be seen from Table 5.4 and Figure 5.24,

Table 5.4: Comparison of the reconstruction errors of the standard PCA, the R-PCA and the FR-PCA for several levels of salt & pepper noise showing RMS reconstruction-error per pixel given by (a) mean and standard deviation and (b) median, lower- and upper quartile.

(a)

Salt&Pepper Noise	Error per Pixel									
	10%		20%		30%		50%		70%	
	mean	std	mean	std	mean	std	mean	std	mean	std
PCA	11.77	5.36	14.80	4.79	18.58	4.80	27.04	5.82	36.08	7.48
R-PCA	11.53	7.18	11.42	7.17	11.56	7.33	11.63	7.48	15.54	10.15
FR-PCA	11.48	6.86	11.30	6.73	11.34	6.72	11.13	6.68	14.82	7.16

(b)

Salt&Pepper Noise	Error per Pixel														
	10%			20%			30%			50%			70%		
	median	Q _{.25}	Q _{.75}	median	Q _{.25}	Q _{.75}	median	Q _{.25}	Q _{.75}	median	Q _{.25}	Q _{.75}	median	Q _{.25}	Q _{.75}
PCA	10.87	7.49	15.30	14.17	11.11	17.66	18.03	15.17	21.25	26.40	23.02	30.17	35.51	30.69	40.09
R-PCA	10.74	6.19	16.28	10.56	6.17	16.25	10.70	6.32	16.27	10.56	6.18	16.23	13.99	7.98	21.03
FR-PCA	10.99	6.37	16.22	10.71	6.37	15.97	10.77	6.47	15.84	10.47	6.31	15.49	14.69	10.11	19.18

all methods can generally cope better with salt & pepper noise. However, also for this experiment FR-PCA yields the best results.

Finally, we evaluated the runtime[†] for the applied different PCA reconstruction methods, which are summarized in Table 5.5. It can be seen that for the given setup compared to R-PCA for a comparable reconstruction quality the robust reconstruction can be speeded up by factor of 18! This drastic speed-up can be explained by the fact that the refinement process is started from a set of data points mainly consisting of inliers. In contrast, in [75] several point sets (hypotheses) have to be created and the iterative procedure has to be run for every set resulting in a poor runtime performance. Reducing the number of hypotheses or the number of initial points would decrease the runtime, but, however, the reconstruction accuracy gets worse. In particular, the runtime of our approach only depends slightly on the number of starting points, thus having nearly constant execution times. Clearly, the runtime depends on the number and size of used eigenvectors. Increasing one of those values, the gap between the runtime for both methods is even getting larger.

[†]The runtime is measured in MATLAB using an Intel Xeon processor running at 3GHz. The resolution of the images is 192x144 pixels.

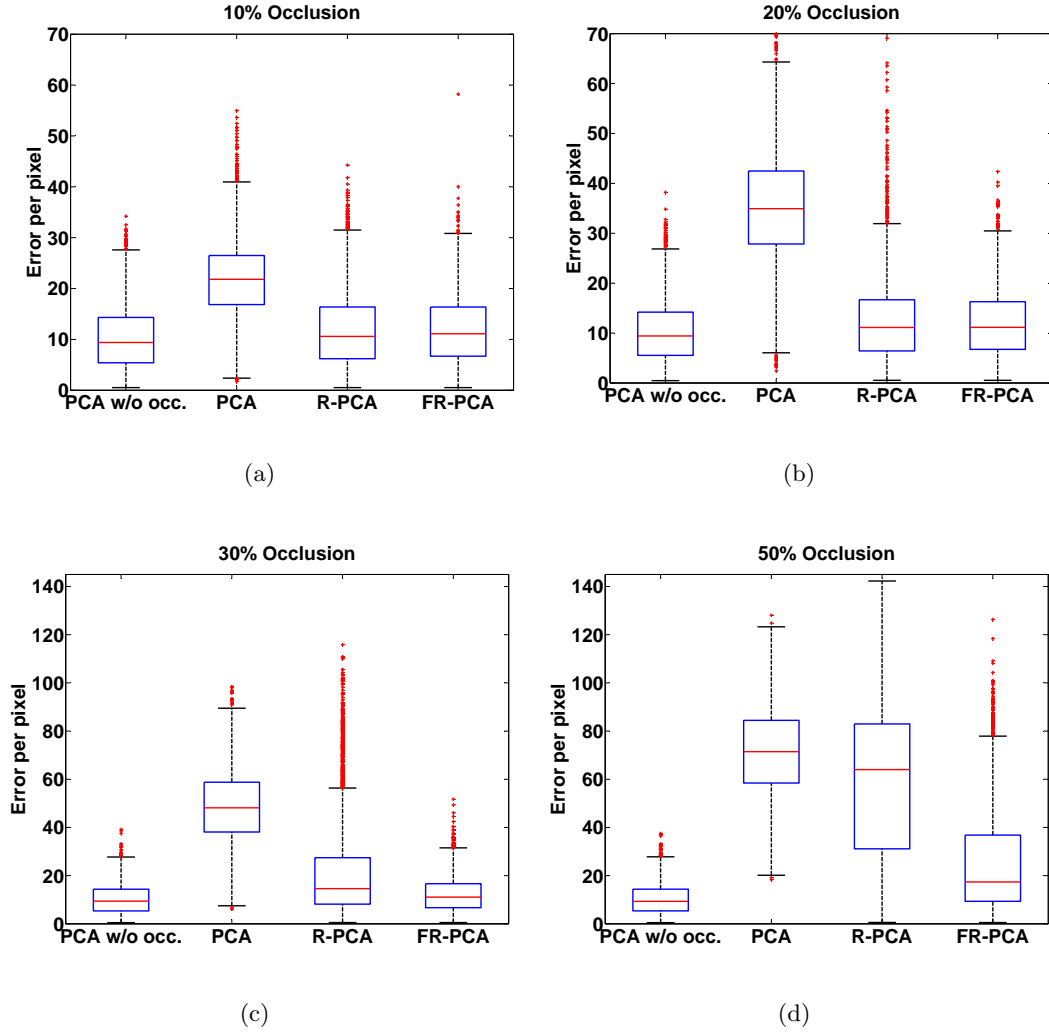


Figure 5.23: Box-plots for different levels of occlusions for the RMS reconstruction-error per pixel. PCA without occlusion is shown in every plot for the comparison of the robust methods to the best feasible reconstruction result.

Table 5.5: Runtime comparison. Compared to R-PCA, FR-PCA speeds-up the computation by a factor of 18.

Occlusion	Mean Runtime [s]					
	0%	10%	20%	30%	50%	70%
PCA	0.006	0.007	0.007	0.007	0.008	0.009
R-PCA	6.333	6.172	5.435	4.945	3.193	2.580
FR-PCA	0.429	0.338	0.329	0.334	0.297	0.307

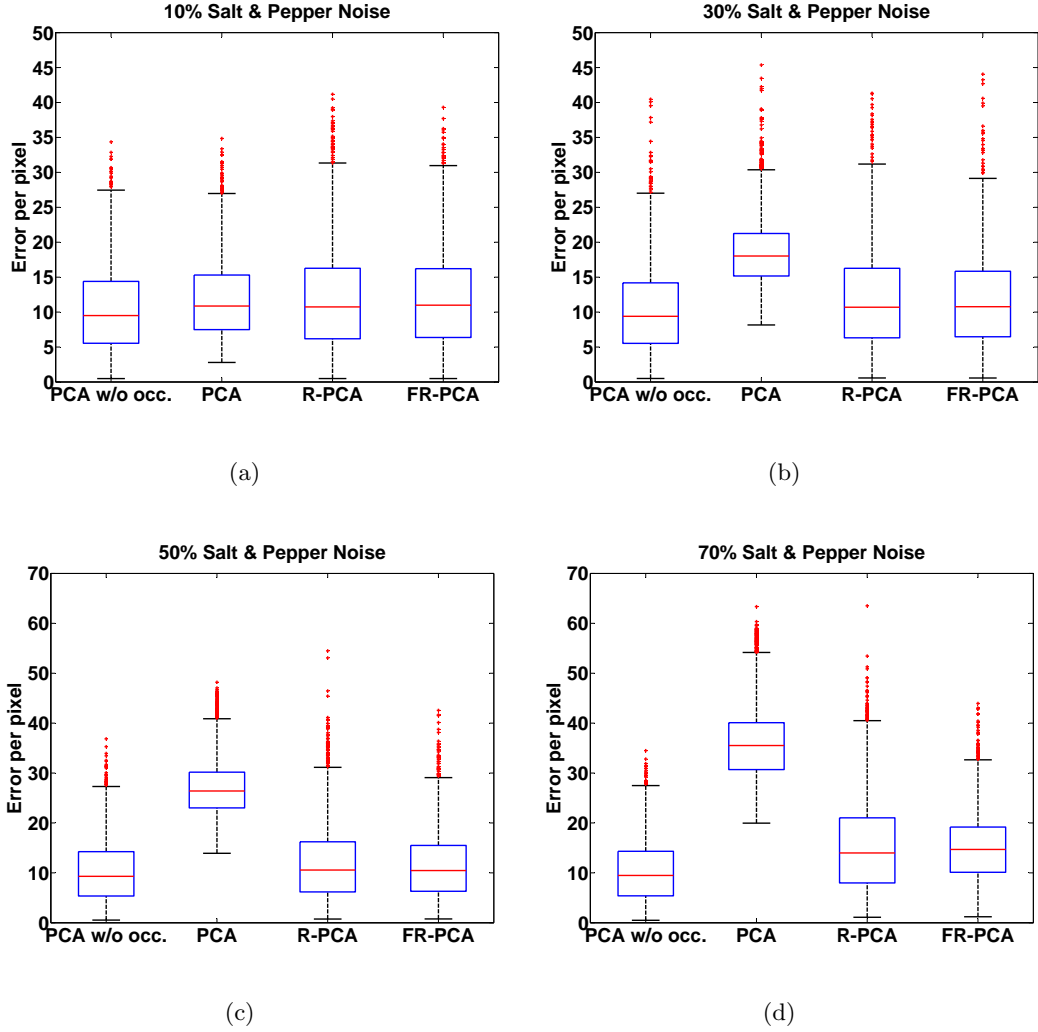


Figure 5.24: Box-plots for different levels of salt & pepper noise for the RMS reconstruction-error per pixel. PCA without occlusion is shown in every plot for the comparison of the robust methods to the best feasible reconstruction result.

5.3.4 Summary

We developed a novel fast robust PCA (FR-PCA) method based on an efficient two-stage outlier detection procedure. The main idea is to estimate a large number of small PCA sub-subspaces from a subset of points in parallel. Thus, for a given test sample, those sub-subspaces with the largest errors are discarded first, which reduce the number of outliers in the input data (gross outlier detection). This set – almost containing inliers – is then used to robustly reconstruct the sample by minimizing the least square reconstruction

error (refinement). Since the gross outlier detection is computationally much cheaper than the refinement, the proposed method drastically decreases the computational effort for the robust reconstruction. In the experiments, we show that our new fast robust PCA approach outperforms existing methods in terms of speed and accuracy. Thus, our algorithm is applicable in practice and can be applied for real-time applications such as robust Active Appearance Model (AAM) fitting. Since our approach is quite general, FR-PCA is not restricted to robust image reconstruction.

Conclusion

The analysis of facial images is not an exact science, but rather subjective depending on biological factors like gender, race, facial hair or age, and extrinsic factors like occlusion, noise, lighting and pose variations. Even for humans it is difficult to interpret facial expressions, especially with different cultural background.

In this thesis we presented a facial analysis system for ICAO compliant facial analysis. The ICAO standard provides a bunch of requirements to assess facial photographs as suitable or improper for biometric applications to prepare automated face verification or identification. Motivated by the needs of our facial analysis system we developed several novel algorithms to address the huge variety of facial appearance and also many extrinsic image factors. We started with a procedure called tokenization to bring arbitrary input images to a standardized coordinate frame. Therefore we need a robust face and facial component detection. All subsequent facial analysis steps rely on the tokenization. The tokenization procedure is very robust when processing near frontal images. If the deviation from the frontal pose is too large, then the tokenization fails. In our case, the system would state that the facial image is not suitable according to the ICAO specification. The cause for the failure could be a large deviation from the frontal pose but it could also be a large occlusion. In that case we cannot state the reason for the non-compliance. But that is ok for our purpose, because we have to determine suitability. To advance the tokenization we could use a robust multi-view face detector. Then it would be easier to determine the reason for non-compliant facial images.

After the detection of the eyes and mouth components, they are assessed if they show an open or closed state. To be robust, we propose to fuse several classifiers to utilize the strengths of the single classifiers. Our algorithm shows very good results, but can be

further improved by adding additional complementary classifiers and fusion schemes.

The ICAO standard prohibits faces differing from the frontal pose. Therefore, we propose three algorithms determining the pose of the head. Still an unsolved issue is the estimation of the pose in case of occluded facial images.

The last criterion we have to deal with is occlusion. We show how to detect occlusions in a facial image. Furthermore, we propose a robust AAM fitting strategy, which is based on our novel Fast-Robust PCA approach. We will think about, how to incorporate the findings of our occlusion handling to our algorithms and to non-robust algorithms found in the literature. An immediate idea is the investigation of how to incorporate the FR-PCA approach directly into the AAM fitting procedure.

All the approaches were evaluated extensively on several databases and compared to state-of-the-art methods. Some parts of our work have also found the application in a commercial product.

6.1 Future Work

All the algorithms are evaluated on several databases. But still, in the future work we intend to perform more experiments to test the methods on additional challenging databases. Furthermore, we want to extend our facial analysis system to check additional ICAO criteria.

Tokenization To advance the tokenization procedure we could use a robust multi-view face detector. In combination with a head pose estimation approach, all the facial images with a large deviation from the frontal pose could be determined.

In the experiments it turned out that the Active Appearance Model (our own implementation and the publicly available version from Stegmann [124]) does not work very satisfactory in terms of accuracy. It often got stuck in a local minimum and thus the fitting was bad. In the future the recently proposed STASM [93] algorithm should be taken under consideration. Experiments showed that the STASM is very robust against varying illumination conditions or partly occluded facial images and showed excellent performance on our data.

Robust similarity measures could be easily incorporated in the congealing algorithm. This may further improve the alignment quality.

Eyes and Mouth Analysis Based on our findings, further work is necessary to evaluate additional, complementary classifiers and fusion schemes to further improve the overall classification results for both criteria.

Analysis of the Deviation from Frontal Pose The algorithms should also be improved in case of an occluded input image. For the 3D Morphable Appearance Model, the range of possible fittings and thereof the limited pose estimation can be extended to multi-view fitting to cover a larger range of head pose. To get rid of the assumption of the linear relationship between the texture residual and the parameter update, further fitting algorithms should be considered, e.g., the inverse compositional algorithm.

We reproduced the published results on the FacePix database using the Biased Manifold Embedding technique. The algorithm works very well on this database, but it turned out that the algorithm is not useful when processing other databases. One direction for future work would be to get rid of the crucial learning step (the learned relationship between the manifold and a new test sample) or to find a way of a direct mapping of a test sample to the manifold.

In the head pose estimation by non-linear regression, we try to learn a relationship between the images, (in our case, the HOG descriptor) and the corresponding annotated angles using a Support Vector Regression. We will also try different regression strategies, like Gaussian Process regression, k-nearest neighbor regression, linear regression and linear ridge regression. First experiments with these regression schemes surprisingly showed a very good performance of k-nearest neighbor regression. The reason might be the dense distribution of different angles in our training database, created by 2D-renderings from faces from a 3D database.

Occlusion Handling We will think about, how to incorporate the findings of our occlusion handling to our algorithms and to non-robust algorithms found in the literature. An immediate idea is the investigation of how to incorporate the FR-PCA approach directly into the AAM fitting procedure.

The occlusion detection algorithm makes some decisions based on thresholds evaluated on a validation dataset. In the future the thresholds should be adapted automatically depending on the test dataset.

The AAM + FR-PCA approach shows slightly worse results compared to the direct fit if there is no occlusion in the image, because of the unavoidable reconstruction-blur resulting from the FR-PCA reconstruction. Thus, the occlusion should be detected first. Only in case of an occlusion, the FR-PCA should be taken as a preprocessing step.



Acronyms

List of Acronyms

3D-MAM	3D Morphable Appearance Model
3DMM	3D Morphable Model
AAM	Active Appearance Model
AdaBoost	Adaptive Boosting
ASM	Active Shape Model
BME	Biased Manifold Embedding
DOF	Degree of Freedom
EER	Equal Error Rate
FAR	False Acceptance Rate
FR-PCA	Fast-Robust PCA
FRR	False Rejection Rate
GPR	Gaussian Process Regression
GRNN	Gaussian Regression Neural Network
HOG	Histogram of Oriented Gradients
HPES	Head Pose Estimation System
HSV	Hue Saturation Value
ICAO	International Civil Aviation Organization
LE	Laplacian Eigenmaps
LGO	Localized Gradient Orientation
LNMF	Local Non-Negative Matrix Factorization

MAP	Maximum a Posteriori
MDL	Minimum Description Length
PCA	Principal Component Analysis
POSIT	Pose from Orthography and Scaling with Iterations
RANSAC	Random Sample Consensus
RMS	Root Mean Squared
ROC	Region of Convergence
SSD	Sum of Squared Differences
SVM	Support Vector Machine
SVR	Support Vector Regression

Active Appearance Model

The Active Appearance Model (AAM) [22] describes the variation in shape and texture of a training set representing an object. From a mean shape and a mean texture, defined in the coordinate frame of the mean shape, the modes of shape and appearance variation are calculated by applying principal component analysis (PCA) to the geometrically and photometrically aligned training examples. Training examples are manually annotated with respect to their shape (commonly defined as corresponding landmark points) and are required to cover the types of variation present in the images one wants to analyze. Formally, from shape representations \mathbf{x} and texture representations \mathbf{g} a statistical model is built given N training examples (i.e., tuples $(\mathbf{x}_n, \mathbf{g}_n)$). The statistical model is based on the interpretation of this representation as a high-dimensional feature vector. A dimensionality reduction is performed by applying PCA to generate the more compact model

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s$$

$$\mathbf{g} \approx \bar{\mathbf{g}} + \mathbf{P}_t \mathbf{b}_t$$

where shapes and textures are represented by the means $\bar{\mathbf{x}}, \bar{\mathbf{g}}$ and the matrices $\mathbf{P}_s, \mathbf{P}_t$ containing the eigenvectors of the training data. \mathbf{b}_s and \mathbf{b}_t are the parameters of this parametric deformable model. By discarding eigenvectors that correspond to small eigenvalues an approximated model is formed. This is feasible since the discarded eigenvectors contribute to the lowest variance in the training data. The AAM model incorporates another dimensionality reduction on the concatenated model parameters $\mathbf{b} = (\mathbf{W}_s \mathbf{b}_s, \mathbf{b}_t)^T$ with \mathbf{W}_s being a weighting matrix to relate shape and texture representations. The com-

bined appearance model $\mathbf{b} \approx \mathbf{P}_a \mathbf{a}$ is finally used to generate synthetic model instances by adjusting the appearance parameters \mathbf{a} .

AAM model fitting makes use of a learned regression model that describes the relationship between parameter updates and texture residual images. Optimization takes place in a gradient descent scheme using the L_2 norm of the intensity differences (between the synthetic model instances and the given test image) as its cost function and the learned regression model for efficiently approximating the Jacobian of the cost function. The parameters of the cost function are the unknown global pose parameters and the combined appearance parameters (which implicitly define shape and texture parameters). A local minimum of the cost function corresponds to a model fitting solution. Since the minimum is local and the parameter space is very large, multi-resolution techniques have to be incorporated and the fitting requires coarse initialization. Despite the multi-resolution approach, the need for proper initialization marks one of the major drawbacks of this method, with the other drawback being its high sensitivity to image occlusions.

Boosting

AdaBoost [43, 120] is a well studied technique for supervised learning tasks. It has recently shown to be tremendously successful in a variety of object localization and classification applications.

In machine learning AdaBoost [43] is a supervised classification technique to establish a complex nonlinear strong classifier

$$H_M(\mathbf{x}) = \frac{\sum_{m=1}^M \alpha_m h_m(\mathbf{x})}{\sum_{m=1}^M \alpha_m}$$

where \mathbf{x} is a pattern to be classified, $h_m(\mathbf{x}) \in \{-1, +1\}$ are the M easily constructible, weak classifiers, $\alpha_m \geq 0$ are the combining coefficients, and $\sum_{m=1}^M \alpha_m$ is a normalizing factor. $H_M(\mathbf{x})$ is real-valued, however classification is done using the signum function $y(\mathbf{x}) = \text{sign}[H_M(\mathbf{x})]$.

The AdaBoost learning algorithm establishes a sequence of best weak classifiers $h_m(\mathbf{x})$ and their corresponding weights α_m . Confronted with N training examples of the form $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where $y_i \in \{-1, +1\}$ are the class labels, a distribution of the training examples is calculated and updated during learning. After each iteration, examples which are harder to separate are investigated to put more emphasis on these examples. It is sufficient that weak classifiers are able to separate a training set better than simple guessing, i.e. it needs a classification rate larger than 50%. The trained weak classifiers determine the features that have to be evaluated in a new image.

In the original object detection framework [142] the weak classifiers are formed from the thresholded responses of simple Haar wavelet like features, see Figure C.1. By calculating Haar features using a so-called integral image representation, it is possible to achieve high

evaluation efficiency. The integral image at position (x, y) is the sum of all pixel values above and on the left (Figure C.2(a)). Computing the sum of pixel values within the specified rectangle, only three basic computations, $P4 - P3 - P2 + P1$ (Figure C.2(b)) are needed. The accuracy comes from the large number of features that are combined by the boosting approach and from the necessity to evaluate features at different scales and locations. Another important advantage of the Viola and Jones approach is their cascaded structure of multiple strong classifiers, which is ideally suited for sliding window processing in object detection. Earlier cascade stages remove a large number of non-object windows, while it is necessary to traverse the whole complex cascade only for a window of the object class.

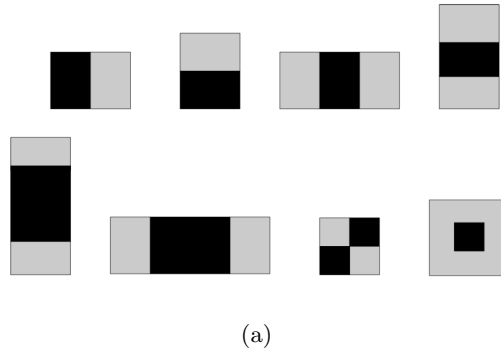


Figure C.1: Haar wavelet-like features [142]

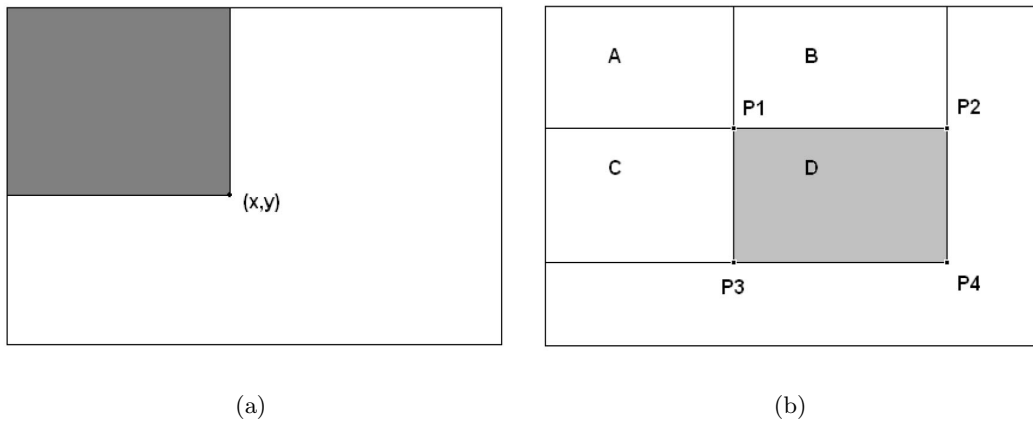


Figure C.2: Integral image representation [142]



Annotation Tool

For our quantitative evaluation of our congealing algorithm, we manually annotated all frontal images of the CVL face database [109]. Therefore we created an annotation tool to annotate facial feature points illustrated in Figure 2.8. The corresponding GUI is shown in Figure D.1. The left pane of the tool is used to select a facial landmark to annotate. The annotated landmark points are displayed in a color coding scheme for easier checking if the landmarks are selected correctly. After the annotation of all visible facial landmarks, the pose of the head can be calculated using a combination of a 3D facial model and the POSIT algorithm [30]. In the right pane, the 3D facial avatar is illustrated after automatically calculating the pose of the head. This avatar can be overlaid on the original input image, see Figure D.2. The whole annotation procedure for a profile face is shown in Figure D.3. After the estimation of the head pose, the points can be back-projected onto the input image, thus showing the quality of the pose estimation (Figure D.3c). When switching to the next image to annotate, the landmark coordinates and the three angles of rotation of the head are saved to a text file. The annotation tool can be controlled with predefined shortcuts to allow an easy usage.

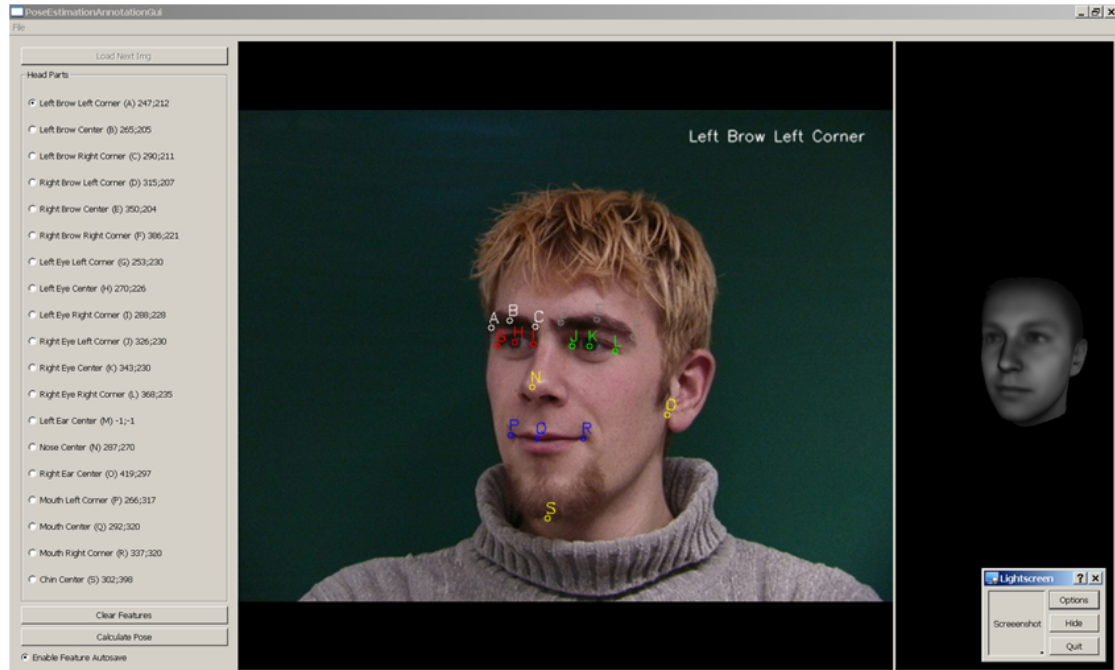


Figure D.1: Screenshot of our annotation tool. The visible facial feature points are annotated.

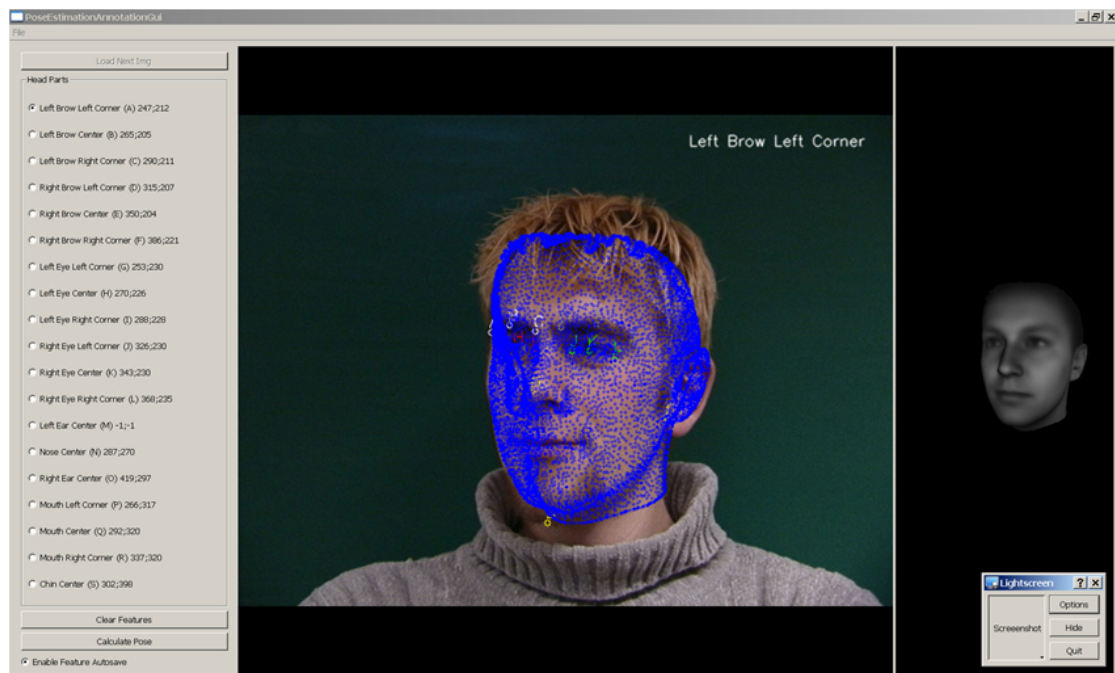


Figure D.2: Screenshot of our annotation tool. In the right pane, the 3D facial avatar is illustrated after automatically calculating the pose of the head. This avatar can be overlaid on the original input image.

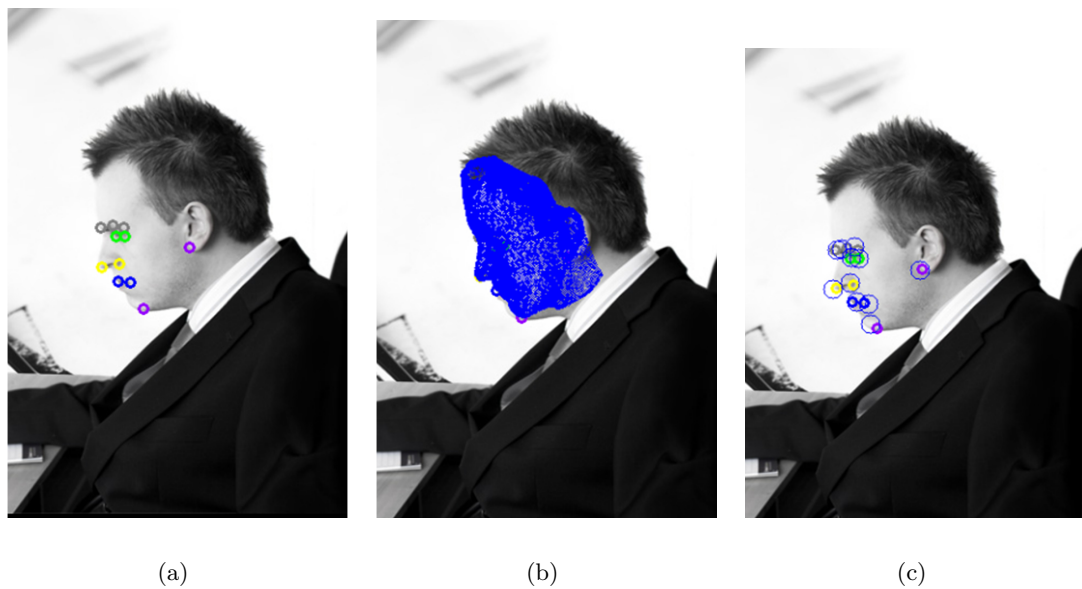


Figure D.3: Annotation procedure. (a) Annotated landmarks, (b) pose estimation, (c) back-projected landmark points

List of Publications

E.1 Book Chapter

M. Storer, P. M. Roth, M. Urschler, H. Bischof, J. A. Birchbauer. Efficient Robust Active Appearance Model Fitting. In *Book Communications in Computer and Information Science (CCIS)*, vol. 68, Springer-Verlag, August 2010

E.2 Peer-Reviewed Conference Papers

M. Storer, M. Urschler, H. Bischof. Intensity-Based Congealing for Unsupervised Joint Image Alignment. In *Proc. 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, August 2010

M. Storer, M. Urschler, H. Bischof. Occlusion Detection for ICAO Compliant Facial Photographs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Workshop on Biometrics, San Francisco, USA, June 2010

M. Straka, M. Urschler, M. Storer, H. Bischof, J. A. Birchbauer. Person Independent Head Pose Estimation by Non-Linear Regression and Manifold Embedding. In *Proc. 34th Workshop of the Austrian Association for Pattern Recognition (AAPR/OAGM)*, Zwettl, Austria, May 2010

M. Storer, M. Urschler, H. Bischof. 3D-MAM: 3D Morphable Appearance Model for Efficient Fine Head Pose Estimation from Still Images. In *Proc. 12th IEEE International Conference on Computer Vision (ICCV)*, Workshop on Subspace Methods, Kyoto, Japan, September 2009

M. Storer, P. M. Roth, M. Urschler, H. Bischof. Fast-Robust PCA. In *Proc. 16th Scandinavian Conf. on Image Analysis (SCIA)*, pp. 430-439, Oslo, Norway, June 2009

M. Urschler, M. Storer, H. Bischof, J. A. Birchbauer. Robust Facial Component Detection for Face Alignment Applications. In *Proc. 33rd Workshop of the Austrian Association for Pattern Recognition (AAPR/OAGM)*, pp. 61-72, Stainz, Austria, May 2009

M. Storer, P. M. Roth, M. Urschler, H. Bischof, J. A. Birchbauer. Active Appearance Model Fitting Under Occlusion Using Fast-Robust PCA. In *Proc. International Conf. on Computer Vision Theory and Applications (VISAPP)*, volume 1, pp. 130-137, Lisboa, Portugal, February 2009

M. Storer, M. Urschler, H. Bischof, J. A. Birchbauer. On Combining Classifiers for Assessing Portrait Image Compliance with ICAO/ISO Standards. In *Proc. Biometrics and Electronic Signatures (BIOSIG)*, volume 137, pp. 153-164, Darmstadt, Germany, September 2008

M. Storer, M. Urschler, H. Bischof, J. A. Birchbauer. Classifier Fusion for Robust ICAO Compliant Face Analysis. In *Proc. IEEE International Conf. on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September 2008

M. Storer, M. Urschler, H. Bischof, J. A. Birchbauer. Face Image Normalization and Expression/Pose Validation for the Analysis of Machine Readable Travel Documents. In *Proc. 32nd Workshop of the Austrian Association for Pattern Recognition (AAPR/OAGM)*, volume 232, pp. 29-39, Linz, Austria, May 2008

Bibliography

- [1] Andrea F. Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, January 2007.
- [2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.
- [3] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.
- [4] S. Baker and T. Kanade. Hallucinating faces. In *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, March 2000.
- [5] Vineeth Nallure Balasubramanian, Sreekar Krishna, and Sethuraman Panchanathan. Person-independent head pose estimation using biased manifold embedding. *Advances in Signal Processing*, 2008:1–15, 2008.
- [6] V.N. Balasubramanian, Ye Jieping, and S. Panchanathan. Biased manifold embedding: A framework for person-independent head pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007.
- [7] R. Beichel, H. Bischof, F. Leberl, and M. Sonka. Robust active appearance models and their application to medical image analysis. *IEEE Trans. on Medical Imaging*, 24(9):1151–1169, September 2005.
- [8] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces versus fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(4):711–720, 1997.
- [9] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computing*, 15(6):1373–1396, 2003.
- [10] D. Beymer. Face recognition under varying pose. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 756–761, 1994.

- [11] Simone Bianco, Francesca Gasparini, and Raimondo Schettini. Combining strategies for white balance. In *Proc. Digital Photography III, IS&T/SPIE Symposium on Electronic Imaging*, 2007.
- [12] J. Black, M. Gargsha, K. Kahol, P. Kuchi, and S. Panchanathan. A framework for performance evaluation of face recognition algorithms. In *Proc. Internet Multimedia Systems II*, July 2002.
- [13] Michael J. Black and Allan D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 329–342, 1996.
- [14] Volker Blanz, Patrick Grother, P. Jonathon Phillips, and Thomas Vetter. Face recognition based on frontal views generated from non-frontal images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 454– 461, June 2005.
- [15] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proc. SIGGRAPH*, pages 187–194, 1999.
- [16] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.
- [17] T. Brick, M. Hunter, and J. Cohn. Get the FACS fast: Automated face analysis benefits from the addition of velocity. In *Proc. Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009.
- [18] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, July 1980.
- [19] Caltech. Caltech face database. <http://www.vision.caltech.edu/html-files/archive.html>, 1999.
- [20] J. Canny. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.
- [21] Chun-Wei Chen and Chieh-Chih Wang. 3D active appearance model for aligning faces in 2D images. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3133–3139, September 2008.

- [22] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [23] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, January 1995.
- [24] T.F. Cootes, K. Walker, and C.J. Taylor. View-based active appearance models. In *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.
- [25] Marius D. Cordea and Emil M. Petriu. A 3-D anthropometric-muscle-based active appearance model. *IEEE Trans. on Instrumentation and Measurement*, 55(1):91–98, February 2006.
- [26] Marius D. Cordea, Emil M. Petriu, and Dorina C. Petriu. Three-dimensional head tracking and facial expression recovery using an anthropometric muscle-based active appearance model. *IEEE Trans. on Instrumentation and Measurement*, 57(8):1578–1588, August 2008.
- [27] Mark Cox, Sridha Sridharan, Simon Lucey, and Jeffrey Cohn. Least squares congealing for unsupervised alignment of images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.
- [28] Mark Cox, Sridha Sridharan, Simon Lucey, and Jeffrey Cohn. Least-squares congealing for large numbers of images. In *Proc. 12th IEEE International Conference on Computer Vision (ICCV)*, August 2009.
- [29] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 886–893, June 2005.
- [30] D. DeMenthon and L.S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15(1-2):123–141, 1995.
- [31] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, 2006.

- [32] F. Dornaika and J. Ahlberg. Face model adaptation using robust matching and active appearance models. In *Proc. IEEE Workshop on Applications of Computer Vision*, 2002.
- [33] Cheng Du and Guangda Su. Eyeglasses removal from facial images. *Pattern Recognition Letters*, 26(14):2215–2220, 2005.
- [34] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, 2001.
- [35] G. J. Edwards, T. F. Cootes, and C. J. Taylor. Advances in active appearance models. In *Proc. 7th IEEE International Conference on Computer Vision (ICCV)*, pages 137–142, 1999.
- [36] J. L. Edwards and J. Murase. Coarse-to-fine adaptive masks for appearance matching of occluded scenes. *Machine Vision and Applications*, 10(5–6):232–242, 1998.
- [37] Hazim Kemal Ekenel and Rainer Stiefelhagen. Why is facial occlusion a challenging problem? In *Proc. International Conference on Advances in Biometrics (ICB)*, 2009.
- [38] P. Ekman and W. Friesen. The facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto, CA, 1978.
- [39] Nikolaos Ersotelos and Feng Dong. Building highly realistic facial modeling and animation: A survey. *The Visual Computer*, 24(1):13–30, 2008.
- [40] V. Erukhimov and K.-C. Lee. A bottom-up framework for robust facial feature detection. In *Proc. 8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.
- [41] Beat Fasel and Juergen Luetttin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003.
- [42] P. F. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.
- [43] Yoav Freund and Robert E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, September 1999.
- [44] Brendan J. Frey and Nebojsa Jojic. Transformed component analysis: Joint estimation of spatial transformations and image components. In *Proc. 7th IEEE*

- International Conference on Computer Vision (ICCV)*, volume 2, pages 1190–1196, 1999.
- [45] Michael Garland and Paul S. Heckbert. Surface simplification using quadric error metrics. In *Proc. 24th Conference on Computer Graphics and Interactive Techniques*, pages 209–216, 1997.
- [46] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam Library of Object Images. *International Journal of Computer Vision*, 61(1):103–112, January 2005.
- [47] Ralph Gross, Iain Matthews, and Simon Baker. Active appearance models with occlusion. *Image and Vision Computing*, 24(6):593–604, 2006.
- [48] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.
- [49] Shu He, John J. Soraghan, and Brian F. O’Reilly. Biomedical image sequence analysis with application to automatic quantitative assessment of facial paralysis. *Journal on Image and Video Processing*, 2007(4):1–11, 2007.
- [50] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 657–662, 2001.
- [51] Erik Hjelm and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, Sep 2001.
- [52] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao. High-performance rotation invariant multiview face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):671–686, April 2007.
- [53] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *Proc. 11th IEEE International Conference on Computer Vision (ICCV)*, 2007.
- [54] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines. In *Proc. 14th International Conference on Pattern Recognition*, pages 154–156, 1998.

- [55] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, 2004.
- [56] Intel. Open computer vision (OpenCV) library. <http://sourceforge.net/projects/opencv>, 2007.
- [57] International Civil Aviation Organization (ICAO). Doc 9303, machine readable travel documents, part 1 - machine readable passport, sixth edition, 2006.
- [58] ISO International Standard ISO/IEC JTC 1/SC37 N506. Biometric data interchange formats - part 5: Face image data, 2004.
- [59] A.K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.
- [60] O. Jesorski, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In J. Bigun and F. Smeraldi, editors, *Audio and Video based Person Authentication (AVBPA)*, pages 90–95, 2001.
- [61] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.
- [62] M. Jones and P. Viola. Fast multi-view face detection. *Technical Report 096*, Mitsubishi Electric Research Laboratories, 2003.
- [63] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, 1(1):137–159, 2008.
- [64] Michael Kirby and Lawrence Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.
- [65] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.
- [66] Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P. W. Pluim. elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. on Medical Imaging*, 29(1):196–205, January 2010.
- [67] Stefan Klein, Marius Staring, and Josien P.W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. *IEEE Trans. on Image Processing*, 16(12):2879–2890, December 2007.

-
- [68] S. Koelstra and M. Pantic. Non-rigid registration using freeform deformations for recognition of facial actions and their temporal dynamics. In *Proc. 8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.
- [69] L. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.
- [70] Fernando De la Torre and Michael J. Black. Robust parameterized component analysis: Theory and applications to 2D facial appearance models. *Computer Vision and Image Understanding*, 91(1-2):53–71, 2003.
- [71] Stephen R. H. Langton, Helen Honeyman, and Emma Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception and Psychophysics*, 66(5):752–771, 2004.
- [72] Erik G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):236–250, 2006.
- [73] Y. LeCun and C. Cortes. The MNIST database. <http://yann.lecun.com/exdb/mnist/>, May 2007.
- [74] Sang-Mook Lee, A. Lynn Abbott, and Philip A. Araman. Dimensionality reduction and clustering on statistical manifolds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [75] Ales Leonardis and Horst Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118, 2000.
- [76] S. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang. Kernel machine based learning for multi-view face detection and pose estimation. In *Proc. 8th IEEE International Conference on Computer Vision (ICCV)*, pages 674–679, 2001.
- [77] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng. Learning spatially localized, part-based representation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 207–212, 2001.
- [78] Stan Z. Li and Anil K. Jain. *Handbook of Face Recognition*. Springer-Verlag New York, 2005.

- [79] Yongmin Li, Shaogang Gong, Jamie Sherrah, and Heather Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, 22(5):413–427, 2004.
- [80] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proc. International Conference on Image Processing*, volume 1, pages 900–903, 2002.
- [81] B. Likar and F. Pernus. A hierarchical approach to elastic registration based on mutual information. *Image and Vision Computing*, 19(1-2):33–44, January 2001.
- [82] Chih-Jen Lin, Ruby C. Weng, and S. Sathya Keerthi. Trust region newton methods for large-scale logistic regression. In *Proc. 24th International Conference on Machine Learning*, pages 561–568. ACM, 2007.
- [83] Dahua Lin and Xiaoou Tang. Quality-driven face occlusion detection and recovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [84] G. Little, S. Krishna, J. Black, and S. Panchanathan. A methodology for evaluating robustness of face recognition algorithms with respect to changes in pose and illumination angle. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 89–92, March 2005.
- [85] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [86] Gareth Loy and Alexander Zelinsky. Fast radial symmetry for detecting points of interest. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(8):959–973, August 2003.
- [87] Yong Ma, Yoshinori Konishi, Koichi Kinoshita, Shihong Lao, and Masato Kawade. Sparse bayesian regression for head pose estimation. In *Proc. 18th International Conference on Pattern Recognition*, pages 507–510, 2006.
- [88] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans. on Medical Imaging*, 16(2):187–198, 1997.

- [89] A. M. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample image per class. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(6):748–763, 2002.
- [90] A.M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, June 1998.
- [91] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, November 2004.
- [92] S. J. McKenna and S. Gong. Real-time face pose estimation. *Real-Time Imaging*, 4(5):333–347, 1998.
- [93] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *Proc. 10th European Conference on Computer Vision (ECCV)*, 2008.
- [94] Steven C. Mitchell, Johan G. Bosch, Boudewijn P. F. Lelieveldt, Rob J. van der Geest, Johan H. C. Reiber, and Milan Sonka. 3-D active appearance models: Segmentation of cardiac MR and ultrasound images. *IEEE Trans. on Medical Imaging*, 21(9):1167–1178, September 2002.
- [95] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):696–710, July 1997.
- [96] H. Moon and M. Miller. Estimating facial pose from a sparse representation. In *Proc. IEEE International Conference on Image Processing*, pages 75–78, 2004.
- [97] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.
- [98] E. Murphy-Chutorian, A. Doshi, and M.M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 709–714, October 2007.
- [99] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. HyHOPE: hybrid head orientation and position estimation for vision-based driver head tracking. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 512–517, June 2008.

- [100] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(4):607–626, April 2009.
- [101] Minh Hoai Nguyen, Jean-François Lalonde, Alexei A. Efros, and Fernando De la Torre. Image-based shaving. *Computer Graphics Forum Journal (Eurographics 2008)*, 27(2):627–635, 2008.
- [102] S. Niyogi and W. Freeman. Example-based head tracking. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 374–378, 1996.
- [103] M. M. Nordstrøm, M. Larsen, J. Sierakowski, and M. B. Stegmann. The IMM face database - an annotated dataset of 240 face images. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2004.
- [104] H. J. Oh, K. M. Lee, and S. U. Lee. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image and Vision Computing*, 26(11):1515–1523, 2008.
- [105] N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. on System Man Cybernetics*, SMC-9(1):62–66, 1979.
- [106] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
- [107] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S. Huang. Human computing and machine understanding of human behavior: A survey. In *Artificial Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Computer Science*, pages 47–71. Springer Berlin / Heidelberg, 2007.
- [108] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, December 2000.
- [109] Peter Peer. CVL face database, Computer Vision Laboratory, University of Ljubljana. <http://www.lrv.fri.uni-lj.si/facedb.html>.

-
- [110] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. Technical report, National Institute of Standards and Technology, Gaithersburg, MD 20899, March 2007.
 - [111] P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.
 - [112] V. S. N. Prasad and B. Yegnanarayana. Finding axes of symmetry from potential fields. *IEEE Trans. on Image Processing*, 13(12):1559–1566, December 2004.
 - [113] Ananth Ranganathan and Ming-Hsuan Yang. Online sparse matrix gaussian process regression and vision applications. In *Proc. 10th European Conference on Computer Vision (ECCV)*, 2008.
 - [114] Rajesh Rao. Dynamic appearance-based recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 540–546, 1997.
 - [115] Elisa Ricci and Jean-Marc Odobez. Learning large margin likelihoods for realtime head pose tracking. In *Proc. International Conference on Image Processing*, 2009.
 - [116] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3D morphable model. In *Proc. 9th IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 59–66, October 2003.
 - [117] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
 - [118] Sam Roweis. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632, 1997.
 - [119] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.
 - [120] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

- [121] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 746–751, 13–15 June 2000.
- [122] Haim Schweitzer. Optimal eigenfeature selection by optimal image registration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 1999.
- [123] Danijel Skočaj, Horst Bischof, and Aleš Leonardis. A robust PCA algorithm for building representations from panoramic images. In *Proc. European Conference on Computer Vision (ECCV)*, volume IV, pages 761–775, 2002.
- [124] M. B. Stegmann. Active appearance models: Theory, extensions and cases. Master’s thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, August 2000.
- [125] Mikkel B. Stegmann and Dorte Pedersen. Bi-temporal 3D active appearance models with applications to unsupervised ejection fraction estimation. In *Proc. International Symposium on Medical Imaging*, volume 5747, pages 336–350, 2005.
- [126] Matthias Straka. Person independent head pose estimation by non-linear regression of histograms of oriented gradients. Master’s thesis, Institute for Computer Graphics and Vision, Graz University of Technology, 2009.
- [127] Matthias Straka, Martin Urschler, Markus Storer, Horst Bischof, and Josef A. Birschbaur. Person independent head pose estimation by non-linear regression and manifold embedding. In *Proc. 34rd Workshop of the Austrian Association for Pattern Recognition (AAPR/OAGM)*, 2010.
- [128] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. Face image validation system. In *Proc. 4th International Symposium on Image and Signal Processing and Analysis*, pages 30–33, 2005.
- [129] Yu-Wing Tai, Michael S. Brown, and Chi-Keung Tang. Robust estimation of texture flow via dense feature sampling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [130] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

- [131] Barry-John Theobald, Iain Matthews, and Simon Baker. Evaluating error functions for robust active appearance models. In *Proc. 7th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 149–154, 2006.
- [132] Philippe Thevenaz, Michel Bierlaire, and Michael Unser. Halton sampling for image registration based on mutual information. *Sampling Theory in Signal and Image Processing*, 7(2):141–171, March 2008.
- [133] Philippe Thevenaz and Michael Unser. Optimization of mutual information for multiresolution image registration. *IEEE Trans. on Image Processing*, 9(12):2083–2099, December 2000.
- [134] Michael E. Tipping and Chris M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 61:611–622, 1999.
- [135] David Tock and Ian Craw. Tracking and measuring drivers’ eyes. In *Proc. 7th British Machine Vision Conference*, volume 14, pages 541–547, August 1996.
- [136] Fernando De la Torre and Michael J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1):117–142, 2003.
- [137] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- [138] USF DARPA Human-ID 3D Face Database. Courtesy of Prof. Sudeep Sarkar, University of South Florida.
- [139] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, 1995.
- [140] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. In *Proc. 7th IEEE International Conference on Humanoid Robots*, 2007.
- [141] Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, 2001.
- [142] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

- [143] Paul Viola and William M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.
- [144] Ying Wang, Kaiqi Huang, and Tieniu Tan. Human activity recognition based on r transform. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [145] J. Weickert, B. M. ter Haar Romeny, and M. A. Viergever. Efficient and Reliable Schemes for Nonlinear Diffusion Filtering. *IEEE Trans. on Image Processing*, 7(3):398–410, 1998.
- [146] Hugh R. Wilson, Frances Wilkinson, Li-Ming Lin, and Maja Castillo. Perception of head orientation. *Vision Research*, 40(5):459–472, 2000.
- [147] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. In L.C. Jain, editor, *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, chapter 11, pages 355–396. CRC Press, 1999.
- [148] B. Wu, H. Ai, C. Huang, and S. Lao. Fast rotation invariant multi-view face detection based on real adaboost. In *Proc. 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 79–84, 2004.
- [149] Bo Wu and Ram Nevatia. Improving part based object detection by unsupervised, online boosting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, June 2007.
- [150] J. Wu and M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138–1158, 2008.
- [151] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade. Real-time combined 2D+3D active appearance models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 535–542, June 2004.
- [152] Lei Xu and Alan L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. on Neural Networks*, 6(1):131–143, 1995.
- [153] Ming Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, January 2002.

-
- [154] P. Yang, Q. Liu, and D. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
 - [155] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3D facial expression database for facial behavior research. In *Proc. 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.
 - [156] D. Yu and T. Sim. Using targeted statistics for face regeneration. In *Proc. International Conference on Automatic Face and Gesture Recognition*, 2008.
 - [157] Xin Yu, Jinwen Tian, and Jian Liu. Active appearance models fitting with occlusion. In *Proc. Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, pages 137–144, 2007.
 - [158] Yin Zhang and Zhi-Hua Zhou. Cost-sensitive face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(10):1758–1769, 2010.
 - [159] Qijun Zhao, D. Zhang, and Hongtao Lu. Supervised LLE in ICA space for facial expression recognition. In *Proc. International Conference on Neural Networks and Brain*, volume 3, pages 1970–1975, Oct. 2005.
 - [160] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.