Invariance and Reliability in Statistical Shape Models

Federico M. Sukno

The cover image design is courtesy of the artist
Milton Hoz de Vila

This thesis was printed with financial support from the Banco Santander grants program at University of Zaragoza and from Scati Labs S.A. (Zaragoza, Spain).

Copyright ©2008 Federico M. Sukno. All rights reserved.
No part of this publication may be reproduced in any form by print, photocopy, digital format or by any other means without prior written permission of the author.

ISBN 978-90-6464-260-9

Printed by Ponsen & Looijen bv, Wageningen, The Netherlands

PHD THESIS

Invariance and Reliability in Statistical Shape Models

UNIVERSIDAD DE ZARAGOZA
Instituto de Investigación en Ingeniería de Aragón

BIOMEDICAL ENGINEERING
Joint PhD Programme from Universidad de Zaragoza and
Universitat Politècnica de Catalunya

Federico M. Sukno

Thesis Director:	**Dr. Alejandro F. Frangi** Universitat Pompeu Fabra, Barcelona, Spain
Reading Committee:	Prof. Vicent Caselles Universitat Pompeu Fabra, Barcelona, Spain
	Dr. José J. Guerrero Universidad de Zaragoza, Zaragoza, Spain
	Prof. Vittorio Murino Università degli Studi di Verona, Verona, Italy
	Prof. Nicolás Pérez de la Blanca Capilla Universidad de Granada, Granada, Spain
	Dr. Jordi Vitrià Universitat Autònoma de Barcelona, Barcelona, Spain

The research described in this thesis was carried out at the research group for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB). Financial support has been provided by a grant from the University of Zaragoza and Banco Santander, and by the Spanish Ministry of Education and Science, Spanish Ministry of Industry and the European Commission through research projects.

Abstract

The present thesis concentrates on the applicability of statistical modeling in facial analysis. Starting from the paradigm of shape and appearance models developed in the last decade, new algorithms are proposed to allow improving their reliability and invariance to different types or rotations. The extensions are formulated in a generic way, such that the models keep the wide applicability of the original approach.

The proposed techniques were experimentally validated in facial analysis tasks. This field became especially relevant in the last few years with an important growth of its international market. A remarkable fact in this sense is the recent adoption of facial biometrics as the standard technology for new biometric passports, taking over other important biometric modalities such as iris or fingerprints. Although the latter are able to achieve lower error rates, the face appearance is the natural way for identification among humans and it is perceived less intrusive. Additionally, it is among the very few biometric modalities that can work, in theory, without the explicit collaboration of the person to be identified.

Chapter 1 provides a brief overview on biometrics and presents the components of a generic biometric system for facial recognition, with especial focus on shape and appearance models. Specifically, Active Shape Models (ASMs) constitute the key methodological component of this thesis, and are briefly covered in Chapter 2.

ASMs allow for the automatic segmentation and analysis of images based on generative models. Introduced in 1992 (by T. Cootes et al.), a considerable number of works has been published on the application of ASMs to diverse types of images, among which medical and facial images are the most numerous. On the other hand, ASMs proofed themselves too simple for modeling purposes in some applications. As a result, later publications focused on extensions and improvements

to the original formulation. One of the most important was the introduction of the Active Appearance Models (AAMs) in 1998. The AAMs soon became popular enough to be considered a separated methodology in its own right, independently from ASMs.

This thesis introduces three novel extensions to ASM. They aim at improving the behavior of these models with a special focus on invariance and reliability. The hypothesis is that the extension and improvement of the segmentation algorithms will lead to a more accurate delineation of facial features, allowing for a more appropriate extraction of image information.

Our first extension addresses the problem of accurate segmentation of prominent features of the face in frontal shots, and is covered in Chapter 3. We propose a method that generalizes linear ASMs using a non-linear intensity model and incorporating a reduced set of differential invariant features as local image descriptors. These features are invariant to rigid transformations, and a subset of them is chosen by Sequential Feature Selection. The new approach overcomes the unimodality and Gaussianity assumptions of classical ASMs regarding the distribution of the intensity values across the training set. Our methodology has demonstrated a significant improvement in segmentation accuracy when compared to the linear ASM, which also derived in lower error rates on identity verification tasks.

The second extension (Chapter 4) concentrates on the invariance of the matching algorithm in the presence of out-of-plane rotations when working with quasi-planar objects. By constraining the analysis to certain parts of the face, the outlines can be approximately considered coplanar. Then, based on projective geometry concepts, ASMs are modified so that they can work independently from the viewpoint (within the range limitations of feature visibility). As a consequence, an ASM constructed with frontal view images can be directly applied to the segmentation of pictures taken from other viewpoints. Validation of the method is presented in images systematically divided into three different rotations (to both sides), as well as upper and lower views due to nodding. The presented tests are among the largest quantitative results reported to date in face segmentation under varying poses.

The third extension (Chapter 5) provides an automatic reliability measure of the segmentation for each analyzed image. That is, the model is able to estimate whether the segmentation obtained for certain image is trustworthy or not. This is very important when ASMs are used into fully-automatic systems, since accurate segmentation is crucial for the subsequent interpretation of the image. The automatic estimation of reliability can be promising for a number of applications. We demonstrate two of them: automatic model selection and reliable identity verification. Results were highly satisfactory in both cases. The strength of the proposed approach relies on its low false positive rate, which means that incorrect segmentations are very unlikely to be misclassified as reliable.

In this way, the first two extensions share the concept of invariance (to rotations in and out of the image plane). On the other hand, it will be shown that the first

extension also increases the accuracy of the segmentation, while the third extension is devoted to the estimation of how reliable is the segmentation of each image. In all cases, intensive experiments have been performed to validate the proposed algorithms, with encouraging results.

Contents

Abstract **5**

Contents **9**

List of Figures **12**

List of Tables **15**

Nomenclature **16**

1 Facial Biometrics **19**
1.1 Biometric passports 20
1.2 A growing market 21
1.3 Challenges ahead 22
1.4 Facial images 24
1.5 Face recognition 27
1.6 Fully automatic systems 28

2 Active Shape Models **31**
2.1 Point distribution model 32
2.2 Appearance model 34
2.3 Matching algorithm 34
2.4 Some relevant extensions 36
2.4.1 Extensions to the matching algorithm 37
2.4.2 Active appearance models 38
2.4.3 Applications 39

2.5 Proposed extensions 40

3 Active Shape Models With Invariant Optimal Features: Application to Facial Analysis 43
3.1 Introduction 44
3.2 Previous approaches 45
3.2.1 Linear ASM 45
3.2.2 Optimal features ASM 46
3.3 Invariant optimal features ASM 46
3.3.1 Irreducible cartesian differential invariants 47
3.3.2 Multi-valued neural network 47
3.3.3 Construction of the intensity model 48
3.3.4 Shape complexities 50
3.3.5 Model matching 50
3.3.6 Feature selection 52
3.4 Experimental evaluation 54
3.4.1 Segmentation accuracy 55
3.4.2 Invariance to image rotations 58
3.4.3 Sub-grid size 59
3.4.4 Feature selection 60
3.4.5 Step by step analysis 62
3.4.6 Identity verification 63
3.5 Discussion 66
3.6 Summary and conclusions 67

4 Projective Active Shape Models for Pose-variant Image Analysis of Quasi-Planar Objects: Application to Facial Analysis 69
4.1 Introduction 70
4.2 Active shape models 73
4.2.1 Point distribution model 73
4.2.2 The intensity model 74
4.2.3 Shape alignment 75
4.3 Projective ASM 75
4.3.1 Projective geometry equations 76
4.3.2 Homography for non-rigid objects 77
4.3.3 Intensity model 78
4.4 Experiments on alignment of point subsets 78
4.4.1 Facial datasets 78
4.4.2 Evaluating point subsets 79
4.4.3 Optimal point subset with exact landmark localization 80
4.4.4 Optimal point subset with uncertain landmark localization . . 81
4.4.5 Conclusions and model initialization 82

4.5 Experiments on segmentation 84
4.5.1 Single-viewpoint model 84
4.5.2 Constraining the transformations 86
4.5.3 Separating the sources of error 88
4.5.4 Multi-viewpoint model 89
4.5.5 Discussion 90
4.5.6 Comparison to related work 92
4.6 Summary and conclusions 96

5 Reliability Estimation for Statistical Shape Models 97
5.1 Introduction 98
5.2 Active shape models 99
5.2.1 Training process 100
5.2.2 Model-to-image adaptation process 100
5.3 Estimating the reliability of the segmentation 101
5.3.1 Defining reliability 102
5.3.2 Estimating reliability from the appearance model 103
5.3.3 Reliability of a whole shape 104
5.3.4 Total probability formulation 104
5.3.5 Conditional probabilities 105
5.3.6 Defining reliability thresholds 106
5.3.7 Incremental accumulation of reliability evidence 107
5.3.8 Summary of the method 108
5.4 Segmentation results 111
5.4.1 Implementational issues 112
5.5 Applications 117
5.5.1 Automatic model selection 117
5.5.2 Reliable identity verification 119
5.6 Summary and conclusions 122
5.7 Appendix: conditional probabilities 124

Bibliography 129

Publications 145

Resumen 147

Acknowledgements 150

Curriculum Vitae 151

List of Figures

1.1 Distribution of the biometrics market by technology in 2006 21
1.2 Annual revenues from biometrics industry between 2005 amd 2010 . . 22
1.3 Typical images from the XM2VTS database 23
1.4 Evolution of the total error rates in to XM2VTS database between 2000 and 2006 24
1.5 Sample images from FRVT 2002 25
1.6 Verification rates for several systems tested within FRVT 2002 against viewpoint changes 26
1.7 Three facial images from the same person showing different expressions 28

2.1 A 98-point template for facial analysis 31
2.2 Shape alignment example 33
2.3 Gray level profiles 35

3.1 Example of sampling with sub-grids 49
3.2 Examples of profiles for OF- and IOF-ASM 51
3.3 Looking for new landmark positions during model matching 51
3.4 ASM, OF-ASM and IOF-ASM segmentation errors per region 57
3.5 ASM and IOF-ASM point-to-curve segmentation errors while varying the regularization constnat of the PDM 58
3.6 Typical segmentation results of ASM and IOF-ASM on images from the AR database 59
3.7 ASM, OF-ASM and IOF-ASM point-to-curve segmentation error for different rotation angles 60

3.8 Segmentation error and time while varying the number of points of the sub-grids . . . 61
3.9 Feature selection statistics from the AR model . . . 63

4.1 Sample image with the corresponding 98-point annotation template . 73
4.2 Effect of non-rigid motion on the estimation of homographies . . . 80
4.3 The 200 best subsets of 4 landmarks to estimate left-right rotations . . 81
4.4 Best points for the estimation of homographies in left-right head rotation . . . 82
4.5 Average error in the estimation of homographies estimated from different number of landmarks . . . 83
4.6 Average error in the estimation of homographies with and without localization error . . . 84
4.7 Best seven points for the initialization of homographies under head rotations and nodding . . . 85
4.8 Segmentation results on AV@CAR dataset grouped by viewpoint . . . 85
4.9 Four iterations of the PASM in a diverging example . . . 86
4.10 Segmentation results with frontal-view models on AV@CAR dataset keeping track of all previous correspondences . . . 87
4.11 Example of a profile view image segmented using models constructed with frontal views only . . . 88
4.12 Median point-to-curve segmentation error on AV@CAR dataset while varying the number of points used to initialize the models . . . 88
4.13 Median point-to-curve errors on AV@CAR dataset using frontal-view models and grouped by viewpoint . . . 90
4.14 Segmentation results from 2-fold cross validation on AV@CAR dataset grouped by viewpoint . . . 91
4.15 First three modes of variation of ASM trained with multiple views . . 92
4.16 First three modes of variation of ASM trained with multiple views . . 92
4.17 Examples of segmentation results for ASM and PASM . . . 93

5.1 Three examples of ASM matches to a face image . . . 99
5.2 Snapshot of the iterative segmentation process for IOF-ASM . . . 101
5.3 Example of reliability estimates and their respective segmentation errors . . . 107
5.4 Example of the segmentation of a facial image from the XM2VTS database . . . 109
5.5 Example of the iterative segmentation process . . . 110
5.6 Segmentation error statistics for the XM2VTS database . . . 115
5.7 Example of the iterative segmentation process considering likelihoods from the third iteration . . . 116
5.8 Sample images of one individual from the AV@CAR dataset . . . 117

5.9 Confusion matrices for ASM and IOF-ASM in the automatic model selection tests . . . 118
5.10 The 49 initial centroids for the segmentation process in the identity verification experiments . . . 121
5.11 Percentage of unreliable segmentations using IOF-ASM for different initializations . . . 122
5.12 Half Total Error Rate using IOF-ASM segmentation for different initial positions . . . 123
5.13 Half Total Error Rate using ASM segmentation for different initial positions . . . 124
5.14 Typical results for the reliable and unreliable segmentations on the multiple-initializations experiments . . . 125

List of Tables

3.1 Tensor and Cartesian formulation of invariants 48
3.2 Composition of the employed datasets and the different groups into which they were divided. 54
3.3 Parameters used to build the statistical models 55
3.4 Point-to-curve segmentation error, normalized to the distance between eye-centers. 56
3.5 Segmentation accuracy and speed for different number of invariants, over 3438 images from AR, Equinox and XM2VTS datasets 62
3.6 Intermediate steps from OF- to IOF-ASM 64
3.7 Identity verification scores using texture parameters 65
3.8 Identity verification scores using shape parameters 65
3.9 Comparison with the segmentation errors reported by other researchers 68

4.1 Comparison of segmentation performance for left and right head rotations with respect to frontal views 94
4.2 Comparison of segmentation performance for left and right head rotations training the models with frontal views 95

5.1 Segmentation Errors [pixels] on AR Database 112
5.2 Segmentation Errors [pixels] on Equinox Database 113
5.3 Segmentation Errors [pixels] on XM2VTS Database 114
5.4 Segmentation errors (average $\pm$ std. error, in pixels) on the AV@CAR dataset with different segmentation strategies 119
5.5 Segmentation errors on the XM2VTS database with 49 different initialization displacements 123

Nomenclature

β	Regularization constant for the Point Distribution Model, page 33
$\mathbf{b}_i$	PCA-space representation of shape *i-th* shape $\mathbf{u}_i$, page 33
b_i^m	*m-th* component of vector $\mathbf{b}_i$, page 33
d_{rr}	A distance metric used to quantify the difference between two homographies, page 79
$\mathcal{E}(\hat{\mathbf{u}}_i)$	Average localization error for shape $\mathbf{u}_i$, page 102
$\epsilon_i^{(j)}$	Localization error for the *j-th* landmark of $\mathbf{u}_i$, page 102
$\mathcal{E}_M$	Threshold for the average localization error of a shape above which the matching process is considered unsuccessful, page 104
$\mathcal{E}_{th}$	Threshold for the average localization error of a shape below which the matching process is considered successful, page 104
$\epsilon_{th}^{(j)}$	Threshold for the localization error of the *j-th* landmark that determines of its positions is considered correct, page 103
$\mathbf{\Phi}$	Eigenvector matrix of $\mathbf{S}$, page 33
$\overline{\mathbf{g}}_j$	Mean intensity profile through the training set for the *j-th* landmark, page 34
$\mathbf{g}_j^i$	Normalized gradient for landmark *j* of the *i-th* training image, page 34
$\mathbf{H}$	A 2D homography matrix, page 77

L Number of landmarks of each shape, page 33

λ_j *j-th* eigenvalue of $\mathbf{S}$, sorted by magnitude in decreasing order, page 33

M Number of eigenvectors kept in $\mathbf{\Phi}$, page 33

N_I Number of images (and shapes) in the training set, page 33

N_{InvAll} In IOF-ASM, all of the available invariant images of the model, page 52

N_{InvF} In IOF-ASM, the number of invariants to be used (after feature selection), page 52

N_T Number of iterations per resolution level, page 103

$\mathbf{\Pi}$ Reference rectangle: bounding box of the facial shape which, if the viewpoint is frontal, becomes a rectangle, page 79

$r_i^{(j)}$ Random binary variable indicating whether landmark j of shape i is correctly placed, page 102

$\hat{r}_i^{(j)}$ Estimation of $r_i^{(j)}$ or reliability estimate provided by the appearance model of the *j-th* landmark based on image evidence, page 102

$R(\hat{\mathbf{u}}_i)$ Reliability estimate for shape $\hat{\mathbf{u}}_i$) based on image evidence, page 104

$\mathbf{S}$ Covariance matrix of the (aligned) shapes in the training set, page 33

$\mathbf{\Sigma}_j$ Covariance matrix for the intensity profile of the *j-th* landmark, page 34

t Current iteration of the model, page 103

$\overline{\mathbf{u}}$ Mean shape of the (aligned) training set, page 33

$\mathbf{U}_i$ Projective representation of shape $\mathbf{u}_i$, page 76

$\mathbf{u}_i$ *i-th* shape, represented as the concatenation of X and Y values in model coordinates (after alignment to the mean), page 32

$\mathbf{V}_i$ Projective representation of shape $\mathbf{v}_i$, page 76

$\hat{\mathbf{u}}_i$ *i-th* shape automatically estimated from the image by the shape model, page 102

$\mathbf{v}_i$ *i-th* shape, represented as the concatenation of X and Y values in image coordinates, page 32

X_G Number of positions sampled for the main grids of IOF-ASM perpendicularly to the profile, page 49

X_g Number of positions sampled for the sub-grids of IOF-ASM perpendicularly to the profile, page 49

$x_i^{(j)}$ Coordinate value for the *j-th* landmark of the *i-th* shape in the X axis, page 33

Y_G Number of positions sampled for the main grids of IOF-ASM in the direction of the profile, page 49

Y_g Number of positions sampled for the sub-grids of IOF-ASM in the direction of the profile, page 49

$y_i^{(j)}$ Coordinate value for the *j-th* landmark of the *i-th* shape in the Y axis, page 33

CHAPTER 1

Facial Biometrics

In a generic definition, biometry (or biometrics) is defined as the science studying measurements and statistics of biological data. In spite of its many possible applications, in the last few years it has been strongly linked to person identification, always based on biological features, as the fingerprint or iris patterns, hand geometry, facial appearance and so on. Indeed, most of the definitions of biometrics refer exclusively to this specific application.

The development of biometrics in the field of security have recently attracted much interest. Several articles have appeared in newspapers and other international media, making the general public well aware about biometrics. One of the driving factors for this (if not the main one) were the concerns of the United States government after the attacks of September 11th (*The New York Times, 24-08-2003*). The need for security reinforcement favored the adoption of biometrics for the control in airports (*El País, 10-10-2002*) and the creation of the new *biometric passports* (*Washington Post, 06-08-2004*), which include face recognition technology to increase the security of travel documents. Both initiatives are currently being adopted by other countries, especially in Europe, and have strongly contributed to the development of biometrics. But among its several modalities, one has been clearly the most favored: facial biometrics (which constitutes the focus of this thesis from the application point of view).

With the goal of increasing airports security there were several tests of face-based identity verification performed since 2001. By 2004 this sort of technology had been also introduced in some casinos, police vehicles and control centers for driving licences (*The New York Times, 31-05-2004*).

However, this technology was not ready to fulfill the enormous expectations suddenly generated. The error rates for facial biometrics in uncontrolled environmental conditions are, even today, too high to enable for massive applications, such as the identity control based on surveillance at an airport. In fact, facial recognition is a relatively recent biometric modality and its performance is known to be lower than other more traditional ones, such as fingerprints (*The New York Times, 05-07-2004*).

On the other hand, when assessing such comparisons, there are two important remarks. First, there is a need for distinguishing between collaborative and non-collaborative scenarios. Fingerprints, iris and hand geometry, for example, require the collaboration of the individual being identified. This is equivalent to a scenario for facial biometrics with highly controlled environmental variables (illumination, viewpoint, expression, etc). Under such conditions the verification rates for face-based systems get easily above 90%. Second, facial biometrics is one of the less intrusive modalities and allows, in theory, to perform the recognition of individuals without requiring their collaboration. Very few technologies compete with faces in this aspect, among which voice and gait are probably the most important. Further, facial biometrics is the only one suitable to be used in surveillance systems (*World Customs Organization, 08-12-2005*).

1.1 Biometric passports

Apart from airports, the other large-scale application of biometrics are the new-generation passports. The United States not only adopted this technology for its own citizens, but has set a deadline for other countries to join the effort. By October 2006, a number of countries whose citizens enjoyed special visa status when traveling to the U.S. must adopt biometric technology on its documents or loose their privileges.

Complying with specifications set by the International Civil Aviation Organization (ICAO), facial biometrics was chosen to be the standard technology for the new passports (*SecureIdNews.com, 19-10-2006*). In the near future, each passport will incorporate a chip on which a facial picture will be stored. Additionally, there would be an option to add biometric data from other modalities (the most probable options being fingerprint or iris).

The decision of using facial biometrics as the core technology for the new documents generated some debate in the United States (*Washington Post, 06-08-2004*). Several experts pointed out that error rates of facial systems were considerably higher than those from other technologies. The main topic of the debate focused on including also fingerprint in the passports. However, the U.S. government delayed fingerprints for future plans, without specifying a date. The choice was based mainly on privacy concerns and negative political effects expected due to the inclu-

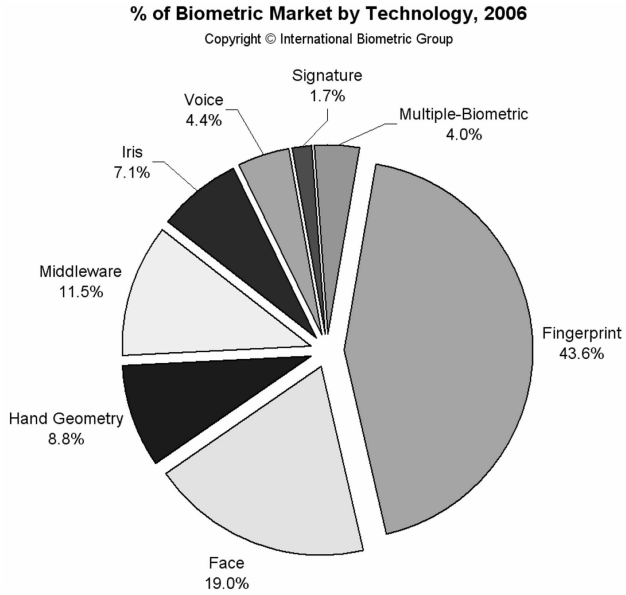

Figure 1.1: Distribution of the biometrics market by technology in 2006, according to a study from the International Biometric Group.

sion of fingerprints. In fact, the final report from ICAO issued in May 2004 stated that facial photographs do not disclose information that the person does not routinely disclose to the public. The use of photographs in travel documents is already socially and culturally accepted worldwide.

1.2 A growing market

In spite of the debate, all passports issued since October 2006 will include a chip with facial information in more than 20 countries (*bbcnews.com, 26-10-2006*). According to official data, only in the United States more than 7 millon new passports are issued every year (*usimmigrationsupport.org, 01-2005*).

In the European Union steps are taking longer, but heading toward the same direction. Already in 2003 Spain, Germany, Italy, France and the United Kingdom approved the inclusion of biometric data into their passports and visas (*El Mundo, 21-10-2003*). In March 2006 the British government started issuing the first biometric passports, with strict requirements for the size and quality of the photographs (*The Independent, 09-04-2006*).

There are also some test programs running at certain airports, aiming at speeding up border controls. In Charles de Golle airport, volunteers choosing to identify

through their fingerprints cross the border in a few seconds, while those choosing the traditional mechanisms usually need 30 to 45 minutes (*El País, 22-10-2006*). Another example is at Heathrow's Terminal 3. Travelers are allowed to enroll into a multi-modal biometric system combining fingerprint, iris and facial images (*The Guardian, 06-12-2006*).

Therefore, it is not surprising that market experts see a great potential for facial biometrics. Especially in the governmental sector and very related to the mandatory adoption of this technology within passports (*www.autoid.frost.com*).

According to a study from the International Biometric Group [75], total revenues from biometrics were about $1200 millions in 2003, with 12% of it corresponding to facial biometircs.. In 2006 this porcentage raised up to 19.2% (Fig. 1.1). And, according to projections from the same study, by 2010 face recognition may get to 21% of the total biometric market, which in turn would have risen more than three times since 2003, reaching $3800 millions (Fig. 1.2).

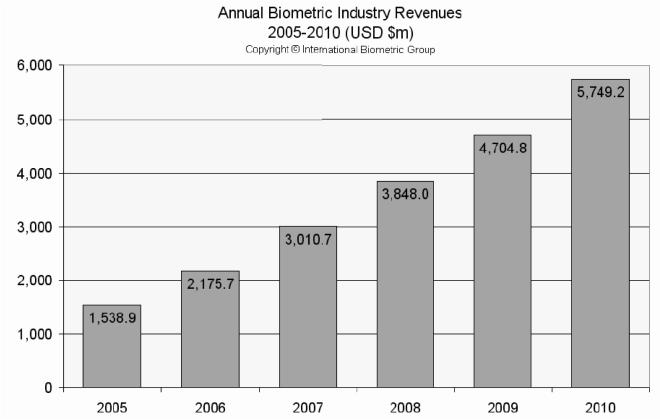

Figure 1.2: Annual revenues from biometrics industry between 2005 amd 2010, according to projections from the International Biometric Group.

1.3 Challenges ahead

The above paragraphs seem to converge into two main ideas: facial biometrics is a promising technology for the near future, but it needs to improve its error rates and the degradations produced by adverse identification scenarios.

Several methods for face recognition were proposed around 1990. Their evaluation in those years was mainly focused on databases containing a few dozens people acquired under strongly controlled conditions. The absence of a standard database hampered for the direct comparison among the different published algorithms.

Some years later, a number of competitions were organized on certain databases with specific protocols. On the verification side, a competition on the XM2VTS database [133] was organized in 2000, and repeated in 2003 and 2006. This database contains 2360 images from 295 people. Fig. 1.3 shows some representative examples while Fig. 1.4 shows the results of the competitions [127,131,132]. When comparing performance in 2006 and 2000, an important reduction of the error rates can be observed. However, the images of this database are too restrictive for a real-scenario application. They were acquired always on a blue background, with an almost perfect frontal pose, neutral expression and uniform illumination (although in 2006 the competition included some lighting variations).

Figure 1.3: Typical images from the XM2VTS database.

On the other hand, the most important competitions on recognition were held between 1994 and 2000 on the FERET database (1994, 1995, 1996 and FRVT 2000 [17, 147, 149–151]). In this case, the number of individuals in the database was progressively increased from 498 people in 1994 to 1196 people since 1996. However, the first large scale evaluation for facial biometrics arrived only in 2002: the Face Recognition Vendor Test (FRVT 2002). This event was based on a database of more than 120,000 pictures from 37,437 people (Fig. 1.5 shows some examples). The data included variations in illumination, expression, viewpoint and short-term aging.

Fig. 1.6 shows the variation of performance for the different systems evaluated against viewpoint change. Almost 100% correct recognition was obtained for frontal shots (as in the experiments over the XM2VTS), but pose changes dropped performance far below 90%, even when using three-dimensional models. Similar plots were constructed for several factors of variation and gathered in the final report of FRVT 2002 [148]. Among the conclusions of that work, some of the most relevant

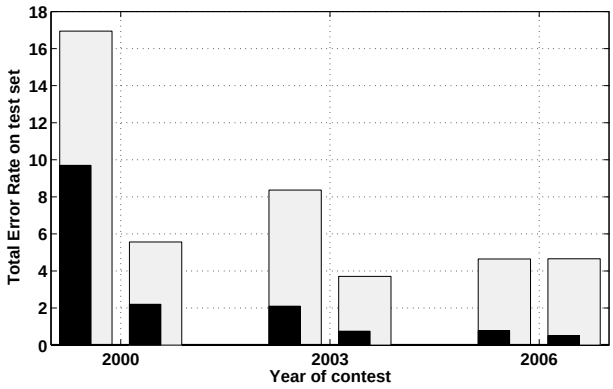

Figure 1.4: Evolution of the total error rates (TER) throughout the three competitions held between 2000 and 2006. The light bars show the average between all presented algorithms while the dark ones show the error rates obtained by the winner. Additionally, the scores are divided into fully automatic systems (left) and semi-automatic systems (right).

are the following:

- Face recognition for indoor images has clearly improved and best current systems are not very sensitive to the normal changes in illumination expected for these environments.
- Performance for outdoor images still needs substantial improvement.
- There is an approximately logarithmic decrease in recognition rates with respect to the elapsed time between enrollment and test images of the same person.
- Recognition rates decrease proportionally to the logarithm of the number of users enrolled in the database.
- Three-dimensional models considerably improve recognition performance in the presence of viewpoint changes.

1.4 Facial images

The appearance of images obtained from any object depend on a series of factors, such as illumination, relative position of the camera (viewpoint), geometric deformations inherent to the object, its albedo and, possibly, partial occlusions. In a study

Figure 1.5: Sample images from FRVT 2002. The upper row shows indoor pictures, while the lower row shows outdoor pictures and includes stronger changes in illumination.

by Moses, Adini and Ullman [3,138] a number of metrics were used to compare the variability of facial images from male subjects (excluding hair, beard and glasses) due to three different factors: illumination, viewpoint and identity. The conclusions showed that identity was (by far) the least important of the three factors. Therefore, a system aiming at perform authentication or identification based on facial images should be robust to the influence of these factors.

A widely used strategy in computer vision is to use prior models of an object. Methods based on this approach are usually classified according to the type of features they extract from the image:

- Methods extracting global features from the images by means of filters applied directly (e.g. Fourier coefficients or moments) or after a thresholding (e.g. number of points in contours or areas). Unfortunately, these techniques are too sensitive to variations in illumination.

- Methods based on edge-properties, which allow for some robustness against illumination changes, since they use relative measures of intensity. The problem in this case is the detection of false edges due to albedo discontinuities, 3D effects and occlusions. On the other hand, the structure of the face does not contain many strong and clearly defined edges. An indicative example is the intrinsic difficulty to precisely define the lower limit of the face in the chin, due to shadows and to the own anatomy.

- Methods based on the detection of key points, such as corners or 3D vertices, where more than two planes of the object intersect. Unfortunately the face

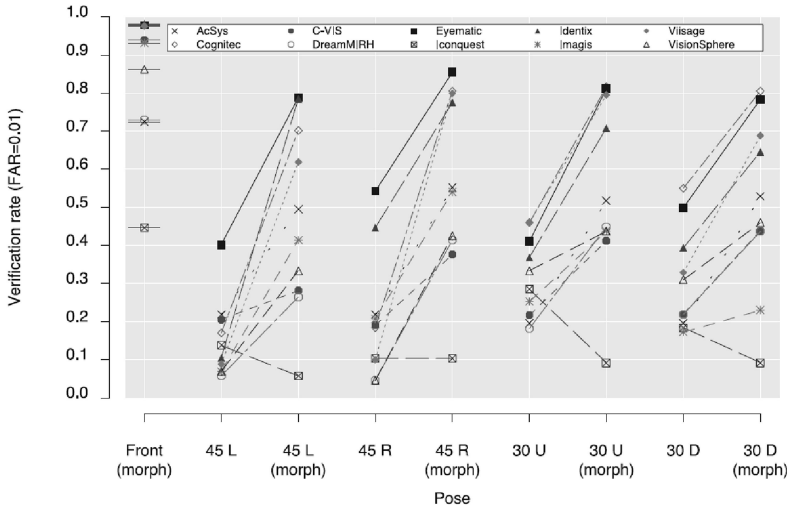

Figure 1.6: Verification rates for several systems tested within FRVT 2002 against viewpoint changes: 45 degrees left and right rotations (45L and 45R) and 30 degrees up and down (30U and 30D). Systems using three-dimensional geometry are indicated as "morph".

provides very few consistent points, such as the corners of the eyes and the mouth.

- Methods based on quasi-parallel boundaries to fit generalized cylinders. Neither this kind of contours are present in the face.

- Methods based on rotational or bilateral symmetry. This property, very common in human-made objects, has been considered approximately true for the human face. Nonetheless, it is only an approximation, as it has been shown by several authors [77, 163].

The most useful properties of the face are associated with certain points from the eyes, mouth, nose, ears and chin. However, even assuming that those points have been found, how shall they be used? While humans easily recognize faces in different situations and in different conditions (even after many years of separation), computer aided face recognition is still one of the most challenging problems in pattern recognition and computer vision [25, 49].

1.5 Face recognition

Since the earliest approaches by Bledsoe [20], Kelly [93], Goldstein [72] or Kanade [89], computer aided face recognition has generated considerable research. Several and diverse algorithms have been proposed. A popular classification of techniques divides them in holistic and feature-based [214]. While holistic methods consider the intensities of the face image as a whole, feature-based methods analyze the image by means of sub-regions accounting for *local features*, such as eyes, mouth, etc. This classification derives from psychological studies, which suggest that both types of analysis are employed in human perception [23, 24]. Indeed, some techniques do combine holistic and feature-based analysis, and they are known as *hybrid approaches*.

The amount of published material on face recognition makes a detailed overview out of the scope of this thesis. The reader is referred to [33, 85, 96, 111, 144, 164, 214] for extensive descriptions of techniques based on computer vision, and to [2,56,123] for others derived from computer graphics and animation. Nonetheless, it is of interest here to consider a different classification of techniques than that provided above, based on the use of statistics. All face recognition algorithms employ some a priori information about the human face. A number of them, especially in early approaches, defined such information *manually* or *ad-hoc*, while others derived it from statistical learning.

In the first category we can find the work by Yuille et al. [209], who modeled facial features from simple lines and arcs. Or the methods based on active contours (or snakes), originally proposed by Kass et al. [92]: certain facial contours were modeled as curves and allowed to deform such that an energy function (depending on image information and regularization terms) was minimized. In yet another work, Sclaroff and Isidoro [168] analyzed tracking of deformable models by means of active blobs so that they can track the head as a rigid cylinder [29].

The main problem with those approaches is the variability of facial images. A human face is a part of a 3D object with no clear boundaries and exhibits an intrinsic variability that is difficult if not impossible to characterize analytically. To overcome these limitations, statistical learning from examples has become very popular. Eigenfaces [95,188], Fisherfaces [11] or Elastic Bunch Graph matching [200] are all examples of this class of algorithms. These methods construct a model of the face based on the statistical analysis of a sample database (the *training set*). The preferred features to be extracted from the images are usually based on texture information (by getting the pixel intensities [188], applying banks of filters [200], etc). Then, subspace projection is used to reduce the dimensionality and/or increase the inter-class separability.

Within statistical models, two outstanding approaches are the Active Shape Models (ASMs) [43, 44, 106] and Active Appearance Models (AAMs) [39, 43], which are covered in the next chapter. Several approaches for facial analysis have been

derived from ASMs and AAMs. For example, the morphable model of Blanz and Vetter [18], which has been one of the most successful algorithms dealing with the 3D nature of the face.

1.6 Fully automatic systems

Over the past 20 years, research has focused on how to make face recognition systems fully automatic by tackling problems such as the localization of a face in a given image or video clip and the extraction of features such as eyes, mouth, etc. [144, 164].

Despite significant advances made in the design of classifiers for successful face recognition, their performance is strongly influenced by the accuracy of the feature extraction. And the latter, at its turn, by the precision of the face detection [48]. Indeed, already in 1971 Goldstein et al. noticed that the most representative properties of the face are associated with points of the eyes, mouth, nose, etc [72]. If these regions are not correctly localized, then the recognition system will probably fail, no matter the strength of the classification algorithm. For example, in the Face Verification Competition on the XM2VTS Database of 2003 [131], the best TER (Total Error Rate [175]) scores for a fully automatic system were 3.86% and 2.10%, for Configurations I and II respectively. On the other hand, when manual registration of the facial images was allowed, almost half of the participant systems beat those scores.

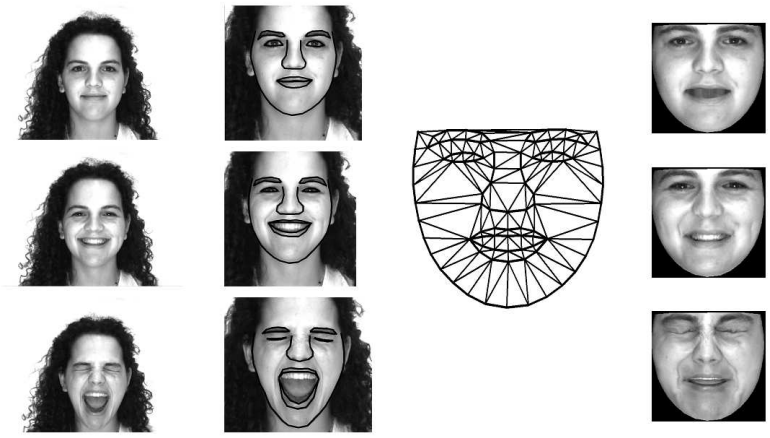

Figure 1.7: Three facial images from the same person showing different expressions (AR database). From left to right and for each row: the original picture, the ASM segmentation, the reference shape with a Delaunay triangulation, and the resulting warping of the segmented image into the reference shape.

A fully automatic face recognition system can be divided into three parts [214]: 1) localization of faces (detection and/or segmentation) from cluttered scenes, 2) feature extraction from the face regions, and 3) recognition or verification. Many algorithms cover more than one step at the same time, or even the three of them. For example, Eigenfaces do benefit from a previous detection step, and then provide the feature extraction and classification, while Elastic Bunch Graph do not require previous detection, as they include it themselves.

The contribution of ASMs and AAMs is mainly focused on steps 1 and 2. These models can accurately provide a segmentation of the prominent facial features, allowing for their comparison based on either shape or image intensity taking into account region correspondences. Fig. 1.7 illustrates a commonly used strategy [106, 204]. When the ASM/AAM algorithm is applied to the photograph of a person (to be identified) it finds a set of contours outlining some predefined facial features (in the example, eyes, brows, lips, nose and silhouette contour). As a result, there is a set of shape parameters for that particular image, which can be used to classify identity [103].

Additionally, the contours are defined as interconnected sets of points. It is possible to use those points as the vertices of a triangular mesh (i.e. by applying a Delaunay triangulation) generating a triangle-to-triangle correspondence of any pair of segmented images. In Fig. 1.7 this has been exploited to map the texture of each face to a *reference shape*. The images on the rightmost column of Fig. 1.7 are usually referred to as *shape-free patches*, as they all share the same shape (the reference), but their textures have been warped triangle by triangle (piecewise warping [162]) from their corresponding segmented images. One important property of shape-free patches is that they allow for the comparison of corresponding textures, irrespective (to a certain extent) of expressions changes and head rotations.

Therefore, the information that can be obtained based on ASM/AAM segmentation can be divided into shape and texture. In fact, the AAM representation is a hybrid combination of shape and texture parameters. Such parameters have been used for identity classification in several works (e.g. see [104, 106, 118, 182, 199, 204]).

The hypothesis motivating this thesis is that the extension and improvement of the segmentation algorithms will derive in a more accurate localization of the facial outlines at the detection/segmentation step of recognition systems, thus improving the extraction of facial features from the image. Therefore, the preferred evaluation metric for the proposed techniques is the segmentation accuracy and its robustness to the influence of degradation factors.

The third step (classification) is beyond the scope of this work. Nonetheless, a suitable implementation of this block has been addressed so that the influence of the proposed algorithms in identification tasks can be also assessed, at acceptable rates within the state of the art.

CHAPTER 2

Active Shape Models

Active Shape Models (ASMs) were introduced by Cootes et al. in 1992 [36]. They constitute a model-based approach in which a priori information of the class of objects to deal with is encoded into a template. Such template is user-defined and allows for the application of ASMs to any class of objects, as long as they can be represented with a fixed topology. Fig. 2.1 illustrates a typical ASM template suitable to represent the facial geometry with 98-points. Other examples using 58-, 64- and 133-point templates can be found in [42, 139, 169].

Figure 2.1: Landmark positions and resulting contours of a 98-point template for facial analysis: 25 points were used for the silhouette outline, 10 for each eyebrow, 8 for each eye, 12 for each lip, and 13 for the nose. The displayed images belong to the AR database and the landmark points have been manually placed.

The template can be thought of as a collection of contours, each defined as the concatenation of certain key points known in the shape analysis literature as *land-*

marks [60]. This constitutes a difference with respect to the Active Contours of Kass et al. [92], which define contours by means of splines. But the more fundamental contribution of ASMs is probably that the deformation allowed in the model template is learnt from a training database. Active contours, on the other hand, constrain the deformation of the model based on functional regularization which penalizes contour properties like smoothness and bending.

As a result, an important property of ASMs is that they are generative models. That is, once trained, ASMs are able to reproduce samples observed in the training database and, additionally, they can generate new instances of the object not present in the database but consistent with the statistics learnt therefrom.

ASMs are based on the combination of a Point Distribution Model (PDM) plus a set of local image appearance models. The PDM describes the shape variability of the template and the appearance models describe the image variability around each of its points. In the following Sections each of the parts is briefly described and the basic equations are provided. For further details the reader is referred to [42, 44].

2.1 Point distribution model

In order to construct the PDM there is the need for a *training set*. The training set consists of a set of images, representative of the object class to be modeled (e.g. faces) that must be annotated with the predefined template. The set of annotated landmarks on one image is usually referred to as *the shape* associated to that image.

The PDM is constructed by applying Principal Component Analysis (PCA) to the set of shapes in the training set. It is generally preceded by a 2D alignment in order to make the analysis independent from 2D rotation and scaling variations. Indeed, *shape* is usually defined as all the geometrical information remaining when positional, scaling and rotational effects have been filtered out from an object [60].

Let $\mathbf{v}_i$ denote the set of landmarks annotated on the *i-th* image on the training set (expressed in *image coordinates*), and $\mathbf{u}_i$ the corresponding shape after similarity alignment:

$$\mathbf{u}_i = s_i \mathbf{R}_i \mathbf{v}_i + \mathbf{t}_i \tag{2.1}$$

Here, $\mathbf{R}_i$ is a rotation matrix, $\mathbf{t}_i$ is the translation vector and s_i is the scale factor, all of them found by aligning every $\mathbf{v}_i$ shape into a common coordinate system: the *model coordinates frame* (see Fig. 2.2). Cootes et al. [44] use Procrustes Analysis [74] for this purpose, minimizing the sum of distances to the mean shape by means of a similarity transformation. At every iteration, each shape is also re-scaled such that $|\mathbf{u}| = 1$.

Assuming N_I images are available, each annotated with L landmarks expressed as x and y coordinates in 2D Euclidean space, the PDM equations can be expressed

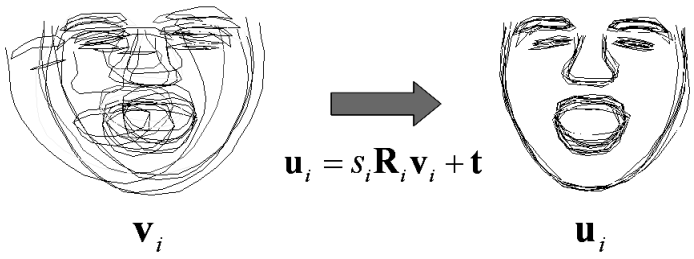

Figure 2.2: Aligning example. The shapes on image coordinates, containing rotation and size variations (left) are aligned into a normalized coordinate system, where only shape variation remains (right).

as follows:

$$\mathbf{u}_i = (x_i^{(1)}, y_i^{(1)}, x_i^{(2)}, y_i^{(2)}, ..., x_i^{(L)}, y_i^{(L)})^T \tag{2.2}$$

$$\overline{\mathbf{u}} = \frac{1}{N_I}\sum_{i=1}^{N_I}\mathbf{u}_i \tag{2.3}$$

$$\mathbf{S} = \frac{1}{N_I - 1}\sum_{i=1}^{N_I}(\mathbf{u}_i - \overline{\mathbf{u}})(\mathbf{u}_i - \overline{\mathbf{u}})^T \tag{2.4}$$

where $\overline{\mathbf{u}}$ is the mean shape and $\mathbf{S}$ the covariance matrix of the set, which is decomposed in its eigenvectors $\mathbf{\Phi}$ and eigenvalues λ_j. Naming $\mathbf{b}_i$ to the PCA-space representation of shape $\mathbf{u}_i$, they are related by:

$$\mathbf{b}_i = \mathbf{\Phi}^T(\mathbf{u}_i - \overline{\mathbf{u}}) \tag{2.5}$$

$$\mathbf{u}_i = \overline{\mathbf{u}} + \mathbf{\Phi}\mathbf{b}_i \tag{2.6}$$

It is possible to use only the first M eigenvectors with the largest eigenvalues:

$$\lambda_j < \lambda_{j+1},\ j = 1, 2 \ldots M-1 \tag{2.7}$$

$$M \leq min(N_I - 1, 2L - 4) \tag{2.8}$$

In that case (2.5) and (2.6) become approximations, with an error depending of the excluded eigenvalues magnitude. Furthermore, each component of the $\mathbf{b}$ vectors is bounded to ensure that only *valid shapes* are represented:

$$|b_i^m| \leq \beta\sqrt{\lambda_m}$$
$$1 \leq i \leq N_I, \quad 1 \leq m \leq M \tag{2.9}$$

where β is a regularization constant, usually set between 1 and 3, according to the degree of flexibility desired in the shape model.

2.2 Appearance model

As stated before, ASMs have as many intensity models as the number of landmarks in the template. Each appearance model is constructed by computing second order statistics for the normalized image gradients, sampled at each side of the landmarks, perpendicularly to the shape's contour, hereinafter, the profile. In other words the profile is a fixed-size vector of values (in this case pixel intensity values) sampled along the perpendicular to the contour such that the contour passes right through the middle of the perpendicular. The appearance model for the *j-th* landmark can be written as:

$$\mathbf{G}_j = \{\overline{\mathbf{g}}_j, \mathbf{\Sigma}_j\} \tag{2.10}$$

where $\overline{\mathbf{g}}_j$ is the mean profile through the training set for the *j-th* landmark, and $\mathbf{\Sigma}_j$ its corresponding covariance matrix. If $\mathbf{g}_j^i$ is the normalized gradient for landmark j of the *i-th* training image:

$$\overline{\mathbf{g}}_j = \frac{1}{N_I}\sum_{i=1}^{N_I}\mathbf{g}_j^i, \quad 1 \leq j \leq L \tag{2.11}$$

$$\mathbf{\Sigma}_j = \frac{1}{N_I - 1}\sum_{i=1}^{N_I}(\mathbf{g}_j^i - \overline{\mathbf{g}}_j)(\mathbf{g}_j^i - \overline{\mathbf{g}}_j)^T \tag{2.12}$$

Fig. 2.3 illustrates a typical gradient computed on the silhouette of the facial contour. The intensity values are sampled perpendicularly to the contour defined by the landmarks. When the gradient is computed, a prominent peak is found at the boundary between face and background.

2.3 Matching algorithm

When the shape models are used for segmentation, only two inputs are required: an image containing a face and a starting guess of the face position (i.e. provided by a face detector). The matching process alternates image driven landmark displacements and statistical shape constraining as in (2.9) based on the PDM, usually performed in a multiresolution fashion in order to extend the capture range of the algorithm. The matching process can be summarized in the following steps:

1 Place a first guess of the model into the image (generally, a scaled version of the meanshape, depending on the application task).

2 Search the image in the neighborhood of each landmark. Adjust the coordinates of each landmark to the best position in this neighborhood. In other words: move the landmarks according to their appearance models. This will generate a cloud of points without shape constraints.

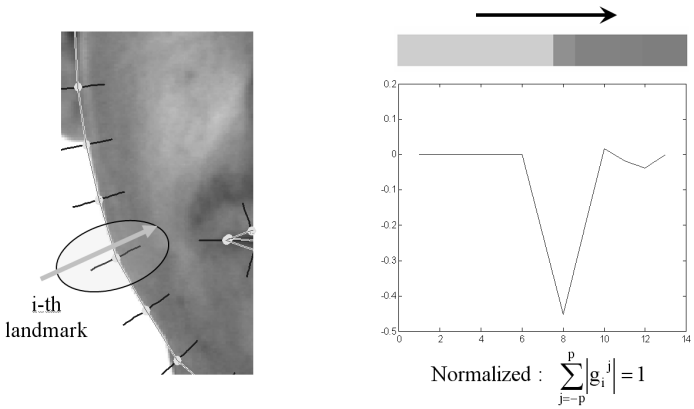

Figure 2.3: On the left, a partial view of a facial image, with the perpendiculars to landmark points indicated. The intensity values sampled for the profile of the highlighted landmark are displayed on the right, as well as the profile resulting after normalization.

3 Apply shape constraints: find the best plausible shape matching the cloud of points generated in step 2. This implies finding the model parameters and some transformation (e.g. a similarity) from model coordinates to image coordinates. The β parameter restricts the PCA coefficients to lie within $\pm\beta$ times the standard deviation observed in the training set.

4 Go back to step 2 until stop condition is reached.

The criterion used to displace the landmarks at step 2 is the minimization of the Mahalanobis distance based to the Gaussian model learnt during training for each appearance model. Let $\{\mathbf{g}_j(1), \mathbf{g}_j(2), \ldots, \mathbf{g}_j(k_P)\}$ be the set of profiles for k_P candidate positions at landmark j. The position suggested by the appearance model will be the one minimizing:

$$Mh^2(\mathbf{g}_j(k), \mathbf{G}_j) = (\mathbf{g}_j(k) - \overline{\mathbf{g}}_j)^T \, \mathbf{\Sigma}_j^{-1} \, (\mathbf{g}_j(k) - \overline{\mathbf{g}}_j) \tag{2.13}$$

for k varying between 1 and k_P. Once all landmarks have been displaced to their best local position, they form a cloud of points which not necessarily describes a plausible shape for the studied object (i.e. a human face).

At step 3, shape restrictions are applied according to (2.9). As a result, landmarks are displaced again to the nearest plausible shape to the candidate points provided by the appearance models (in a least squares sense). The rational behind shape restrictions is the assumption that facial shapes lie approximately within a hyperellipsoid (in PCA-space) that can be learnt during training. However, for simplicity reasons it is very common to use (2.9) and limit the shape-space to a hyper-cuboid.

2.4 Some relevant extensions

Apart from the original formulation of ASMs introduced above, a considerable number of extensions has been proposed. In this section we briefly overview the most relevant ones, especially emphasizing those related to facial analysis.

One of the most interesting aspects of the original formulation of ASMs is its simplicity. For example, the residuals of the shapes with respect to the mean are assumed Gaussian. This approach works well for a wide variety of examples, although it is too simple to represent non-linear shape variations, such as those generated when there are changes in viewing position of a 3D object. A more general approach is to model the distribution of the residuals using a mixture of Gaussians, as proposed by Cootes et al. [43].

Non-linear formulations of the PDM were proposed by Sozou et al. [174] using Multi-Layer Perceptrons (MLP) to perform the PCA decomposition, similarly to [98]. The experiments reported on image search revealed comparable performance to the linear PDM, yet requiring half the number of dimensions [63].

A non-linear PDM was also proposed by Romdhani et al. [158] to cope with head rotations out of the image plane. They used Kernel PCA to handle the non-linearity, adding the view angle as an additional parameter to the landmarks vector. The application to facial images under multiple viewpoints is probably the most successful for non-linear PDMs and has been also extended to appearance models. For further details, please refer to Chapter 4, which concentrates on multi-view images.

Chen et al. [34] focused on a different aspect of the PDM. They decomposed the overall error of ASM fitting into two terms: representation error and search error. They analyzed the behavior of the error as a function of the variance explained by the model. Based on experiments over 400 faces they claim that the optimal percentage of variance retained by the model is lower than that generally employed.

In yet another research line on PDMs, Rogers et al. [155] addressed the parameter estimation when the PDM must be adjusted to new points. They compared the original formulation (least sum of squares) with robust estimation algorithms, such as random sample consensus (RANSAC) [66] and least median of squares (LMedS) [161]. Experiments synthetically simulating the presence of outliers indicated that robust methods can considerably improve ASM search, but their results on real data were not too encouraging. A similar idea was proposed by Lekadir et al. [108], but with better results. Putting aside that they worked on totally different data, the key contribution of [108] is the use of the ratio of interlandmark distances to detect outliers. This ratio is invariant to the estimation of the pose parameters, which allows for decoupling the outlier analysis and the ASM fitting. Wang et al. [198] also revised the fitting of the PDM, but based on ad hoc weighting of different facial landmarks. They pre-localized some salient features of the face and used them as priors to constrain the PDM optimization.

As opposed to those modifying the PDM, a number of authors have focused on the appearance model of ASMs. Wang et al. combined the first order derivatives with edge information to work with facial images [198] and Koschan et al. explored including color information [97]. However, most of the research in this topic has concentrated in replacing the gray-level profiles with other features. Quantitative comparisons of some simple features based on row image intensities and gradients can be found in the works by Behiels et al. [10] and, more recently, by Seise et al. [170]. More elaborated features include sets of image derivatives [191], Gabor wavelets [87, 183], normalized patches [51], haar-like features [211, 212] and, in a recent work [90], Scale Invariant Image Transform (SIFT) [121].

An interesting point on the work by Ginneken et al. [191] is that the choice of the features has a theoretical justification. It is known from the Taylor series expansion that a function can be approximated (in a neighborhood) by appropriately combining its partial derivatives, up to a given order. Building on this, in Chapter 5 we will present further extensions to the method and demonstrate strong accuracy improvements on the segmentation of facial images.

2.4.1 Extensions to the matching algorithm

Apart from the approaches mainly modifying the PDM or the appearance model of landmarks, several authors have studied the matching strategy. Aiming at improving the efficiency of ASM search and at increasing the initial displacement that can be handled, Cootes et al. proposed a multi-resolution approach [45]. It consists in downsampling the original image to get lower resolution versions of it and then perform the search in a course-to-fine strategy. The gray levels around each landmark are modeled separately for each resolution and the approach was showed to improve the accuracy and speed of the segmentation process. Indeed, most people using ASMs employ this multi-resolution strategy.

A step forward into this line was proposed by Liu et al. [117] with their Hierarchical Shape Model (HSM) by defining a course-to-fine search strategy also for the PDM. The number of landmarks is therefore increased every time the resolution is refined. Additionally, they dismiss the independence assumption for the local image search of landmarks and propose conditional independence with respect to a hidden variable. Global convergence is then addressed by a data driven Markov Chain Monte Carlo method [187]. Similar strategies were proposed by Davatzikos et al. [52] who used the wavelet transform to downsample the shape points, and by Zhang et al. [211] based on the Adaboost algorithm and Haar-like features [194]. Additionally, the approaches from Zhang et al. and Liu et al. integrate face detection within the active shape model, although no quantification was provided in such task. Adaboost is also used by Yan et al. [205] in combination with an iterative optimization algorithm based on a Bayessian framework.

Other authors have argued about the independence of the local image search

for each landmark, indicating that it is not convenient. Proposed alternatives include simple smoothness constraints between neighboring points [135] or low level modeling of shape by means of a Markov Network, which provides additional constraints to those imposed by the limitation of PCA parameters at a higher level [115].

Another group of works explored shape priors in a Bayesian framework to enhance the computation of model parameters through MAP estimation [204,215,216]. Kanaujia et al. [90] combined this approach with a non-linear representation of the shape manifold. The manifold was modeled as the combination of overlapping linear subspaces, which are determined by clustering the training shapes (after alignment to the tangent space). The number of clusters is automatically chosen by employing normality statistics of the cluster distributions. Also within the Bayesian assumption, Tamminen et al. [183] proposed a sequential matching strategy, in which the information about previously identified features is used as prior for the matching of the following ones.

2.4.2 Active appearance models

The application of ASMs in classification problems suggested that the information that may be extracted from image intensities (appearance) confined by the automatic segmentation may be more useful than the shape itself [105]. Building on this idea, Edwards et al. [63] concatenated shape and appearance parameters into a single vector and combined them by means of PCA. As a result, they obtained a new set of *combined* parameters encoding both shape and appearance variation at the same time. This approach inspired the most successful extension to ASMs: the Active Appearance Model (AAM) [37,39], which generated itself a considerable number of extensions.

Yan et al. [207] proposed a Texture Constrained ASM (TC-ASM) which borrows global texture from AAM to constrain the optimization of shape parameters: the conditional distribution of a shape given its associated texture is modeled as a Gaussian. Texture constrains for ASM were also used by Li et al. in [114] but in this case with the goal of better initialization.

Sung et al. [180] combined ASM and AAM by converting the original ASM into a gradient-based optimization problem. They proposed a combined optimization function and compared it to ASM, AAM and TC-ASM, The proposed approach showed the lowest errors but experiments lack in two important aspects: only 80 images were used and several parameters were empirically set according to results on the test data.

One important issue of AAMs is the dimensionality of the texture patch, especially when working at high spatial resolution. The obvious solution of pixel-wise subsampling was shown to produce poor results even at low decimation rates like 4:1 [38]. A solution based on Haar-wavelets was proposed by Wolstenholme et al. [201] that allowed for a dimensionality reduction up to 20:1. Larsen et al. [107]

performed further evaluation of this technique and extended it by incorporating wedgelet representation of the texture. Reduction rates up to 200:1 were reported for wedgelet AAM without excessive drop in segmentation accuracy. However, the wavelet AAM was shown more appropriate for realistic texture reconstruction.

In a recent work, Liu et al. [119] addressed the study of resolution in AAM from a different perspective, focusing on the performance that can be achieved for low resolution images. They concluded that a mismatch between the resolution of the model and the images being fitted can seriously compromise performance. An experiment on dynamic selection of the model resolution was also provided, although under a very simple approach. In another work, Dedeoglu et al. [57] propose to include the image formation process in the appearance model and quantify the degradation produced by low resolution images. They compared the proposed strategy to a single-resolution AAM by upsampling the image to match model's resolution, which does not seem appropriate. The opposite approach would be more plausible (downsampling the model) and involves assumptions on the image formation process as well.

2.4.3 Applications

Apart from facial images, ASMs and AAMs have been used for the segmentation and/or analysis of several types of objects. For example, PDMs have been applied to recovering upper-body pose from silhouettes or skin-colored blobs [136,140], ASMs have been used to locate and model electronic components in printed circuit boards [41], for the automatic segmentation of hand radiographs [173,184], capilaries imaged by electron microscopy [154], thrombus in abdominal aortic aneurysms [54], Magnetic Resonance (MR) images of the carotid artery [108], the heart [190]m the brain [52,62] or the cortical sulci [30]; to find vertebral structures from dual X-ray absorptiometry [172] and to outline several bone structures [10].

The list, far from being complete, illustrates the wide variety of objects that can be analyzed through ASMs, and also AAMs. For example, an interesting application was proposed by Kurt et al. [101] who used AAMs as part of a composite face generation system. Such systems are used by the police to help witnesses draw the face of a suspect [69,71]. In the proposed strategy, an AAM model was built from a face database, so that both shape and appearance were represented by means of the AAM parameter vector. Then, those vectors were used as chromosomes for nature-inspired heuristics, like genetic algorithms.

Although this thesis is concerned exclusively with bi-dimensional (2D) shape models, several authors have proposed to extend the analysis to more dimensions. Most of them have addressed the construction of three-dimensional (3D) models. In the context of facial analysis, the seminal work by Blanz et al. [18] is by far the most important. They used a large dataset of 3D face scans containing both geometric and textural information to construct a 3D face model coined multidimensional

morphing function. This approach has been followed in a number of papers reporting results on 3D face recognition [19, 31, 83, 148, 157]. There are also several works using 3D shape models in medical imaging applications [30, 55, 108, 141, 190], and others at half way between 2D and 3D, which derive 3D shape models from multiple 2D views but perform the image search in 2D [113, 128, 202].

In analogy to the extension to 3D, some researchers have addressed the modeling of shape dynamics, by considering the temporal dimension. In this case, one of the most cited approaches is the one by Bosch et al. [22], who extended Cootes's AAM to time sequences by considering the whole image sequence as a single shape/intensity sample, hence jointly modeling space and time. Similar approaches were followed Hamarneh et al. [79] and Mitchel et al. [134], while Perperidis et al. [146] and Hoogendoorn et al. [81] presented methods to separate the variations in space and time, yet including them within the same statistical model.

2.5 Proposed extensions

One of the first application fields of ASMs and AAMs was facial analysis. The statistical nature of these models results very attractive when dealing with complicated objects, as it is the case of the human face. Some outstanding advantages of ASMs and AAMs can be summarized as follows [42]:

- They constitute a *top-down* approach, in which the construction of a global model ensures the coherence of the result when combining the evidence from different (low-level) regions of the object.
- They are *very flexible* regarding the choice of the representation. As long as there is a fixed topology, the model template can be chosen by the user specifically for the desired application. For example, in [139, 169, 177] facial images were annotated with 58-, 64- and 98-point templates, respectively, yet in all cases outlining silhouette contour, eyes, brows, mouth and nose. Cootes et al. [42] decided to include also hair shape and ears, ending up with a 133-point template.
- There is no need to explicitly regularize the model deformation. Instead, it is statistically learnt from the annotated data, both for the shape and the texture. For example, the allowed variation for facial shapes will never allow for the two eyes to meet together nor to cross each other. On the other hand, as long as the training database includes the appropriate examples, they will be allowed to be opened or closed, even independently.

These facts and the simplicity of the seminal approaches by Cootes et al. [37, 44] have made ASMs and AAMs very popular. However, when this models are used for facial analysis, a number of practical problems decrease their performance and

limit their application. In the context of this thesis, there is a special interest in three of them:

1. *Insufficient accuracy of the segmentation* due to the complexity of the underlying distribution of image intensities in facial images. Indeed, one of the few assumptions made by ASMs is that the normalized gradients for each landmark follow a Gaussian (and unimodal) distribution (see 2.10). This reduces the localization precision of the landmarks (and hence of the facial outlines). As a result, facial features are not accurately extracted, which drops identification performance (Chapter 3).

2. *Limited viewpoint coverage* can be handled, especially when dealing with lateral views of the face (left and right head rotations). It has been experimentally shown that when such variation are greater than ± 20 degrees (± 40 according to other authors) they are difficult to handle by a single model. This is a strong limitation for a facial-based biometric system, especially when they aim at being used in non-collaborative environments, such as in video surveillance applications (Chapter 4).

3. *Absence of reliability measure* for the result of processing a given image (e.g. there is not a trustworthy criterion for convergence). No matter whether the segmentation successfully matches the outlines of a face or it diverged far from it, the result obtained by ASM is a set of points describing a perfectly plausible facial geometry. This situation can be easily discriminated by a human operator, but not by the model, and hampers the implementation of ASMs as a part of a fully automatic system (Chapter 5).

The extensions proposed in this work aim at improving the performance of ASMs when dealing with these three problems. They have been conceived in an application-driven manner, putting the emphasis in their usefulness for facial analysis. However, they were also stated in a generic way, avoiding task-specific assumptions as much as possible.

Thus, the solution proposed for segmentation accuracy can be applied to any object (as ASMs do), although the algorithm has been especially designed to handle object's templates with multiple embedded shape, as those present in facial templates. The same applies to the proposed reliability estimation.

To allow ASMs to work in front of a broader range of image viewpoints, our projective extension does require a special assumption: the points of the template should be coplanar. While this is not the case of a human face as a whole, an important part of it approximately fulfills this constraint. Any other object suitable to be described by an approximately coplanar template may benefit as well from this approach.

In the following three chapters, each of the proposed extensions is explained in detail, including a more thorough discussion about their motivation and related

work, as well as the experimental evaluation.

CHAPTER 3

Active Shape Models With Invariant Optimal Features: Application to Facial Analysis

Abstract - *This work is framed in the field of statistical face analysis. In particular, the problem of accurate segmentation of prominent features of the face in frontal view images is addressed. We propose a method that generalizes linear Active Shape Models (ASMs), which have already been used for this task. The technique is built upon the development of a non-linear intensity model, incorporating a reduced set of differential invariant features as local image descriptors. These features are invariant to rigid transformations, and a subset of them is chosen by Sequential Feature Selection for each landmark and resolution level. The new approach overcomes the unimodality and Gaussianity assumptions of classical ASMs regarding the distribution of the intensity values across the training set. Our methodology has demonstrated a significant improvement in segmentation precision as compared to the linear ASM and Optimal Features ASM (a non-linear extension of the pioneer algorithm) in the tests performed on AR, XM2VTS and EQUINOX databases.*

Adapted from F.M. Sukno, S. Ordas, C. Butakoff, S. Cruz, and A.F. Frangi. Active shape models with invariant optimal features: Application to facial analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1105-1117, 2007.

3.1 Introduction

In many automatic systems for face analysis, following the stage of face detection and localization and before face recognition is performed, prominent facial features must be extracted. This process currently occupies a large area within computer vision research.

A human face is part of a smooth 3D object mostly without sharp boundaries. It exhibits an intrinsic variability (due to identity, gender, age, hairstyle and facial expressions) that is difficult if not impossible to characterize analytically. Artifacts such us make-up, jewelery and glasses cause further variation. In addition to all these factors, the observer's viewpoint (in-plane or in-depth rotation of the face), the imaging system, the illumination sources and other objects present in the scene, may affect the overall appearance. All these intrinsic and extrinsic variations make the segmentation task difficult and hamper a search for fixed patterns in facial images. To overcome these limitations, statistical learning from examples is becoming popular in order to characterize, model and segment prominent features of the face.

An Active Shape Model (ASM) is a flexible methodology that has been used for the segmentation of a wide range of objects, including facial features, e.g. [106]. In the seminal approach of Cootes et al. [44] shape statistics were computed from a training set of shapes and local grey-level profiles (normalized first order derivatives) were used to capture the local intensity variations at each landmark point. In [37] Cootes et al. introduced another powerful approach to deformable template models, namely the Active Appearance Model (AAM). In AAMs a combined PCA of the landmarks and pixel values inside the object is performed. The AAM handles a full model of appearance, which represents both shape and texture variation.

While AAM has its own benefits, like the ability to model texture variation, ASM is fast, mainly due to the simplicity of its texture model. The latter is constructed with just a few pixels around each landmark whose distribution is assumed to be Gaussian and unimodal. This simplicity, however, turns into weakness when complex textures must be analyzed. In practice, local grey-levels around the landmarks can have large variations and pixel profiles around an object boundary are not very different from those in other parts of the image. To provide a more elaborated intensity model, van Ginneken et al. [191] proposed the Optimal Features ASM (OF-ASM). It is non-linear and allows for multi-modal distribution of intensities, since it uses k-nearest neighbor (kNN) classification of local texture descriptors (*jets*), based on image derivatives. Wiskott et al. [200] also use local jets to construct their Elastic Bunch Graph. The latter are based on Gabor kernels and constrain the shape variations by an elastic model rather than by a statistical point distribution model as in ASMs. Although this method is mainly oriented towards an in-class recognition task, it can be also applied to obtain segmentations of facial images, achieving satisfactory results.

The main contribution of the OF-ASM was an increased accuracy in the seg-

mentation task, which has shown to be particularly useful in segmenting objects with textured boundaries in medical images. However, its application to facial images is not straightforward: facial images have a more complex geometry of embedded shapes and present large texture variations for the same *region* across different individuals. In this work we will discuss those problems and develop modifications to the model in order to make it deal with facial complexities. The OF-ASM derivatives will also be replaced so that the intensity model is invariant to rigid transformations. The new method, coined Invariant Optimal Features ASM (IOF-ASM) [178] also tackles the problem of the segmentation speed, which was a drawback in OF-ASM [191]. It will be shown that IOF-ASM offers the possibility to trade off between segmentation accuracy and speed. The performance of our method will be compared against both the original ASM and the OF-ASM, using the AR [126], XM2VTS [133] and Equinox [171] databases as test beds. Experiments were split into segmentation accuracy and identity verification tests, based on the Lausanne protocol [133].

The remainding of this paper is organized as follows: in Section 3.2 we briefly describe the underlying theory of ASM and OF-ASM approaches, while in Section 3.3 the proposed IOF-ASM is presented. In Section 3.4 we describe the materials and methods for the evaluation and show the results of our experiments, which are discussed in Section 3.5. Section 3.6 summarizes and concludes the paper.

3.2 Previous approaches

3.2.1 Linear ASM

In its original form [44], ASM is built from sets of prominent points known as *landmarks* [44] by computing a Point Distribution Model (PDM) and a local image intensity model around each of those points.

The PDM is constructed by applying PCA to the aligned set of shapes, each represented by landmarks. The original shapes $\mathbf{u}_i$ and their model representation $\mathbf{b_i}$ ($i = 1, ..., N_I$) are related by the mean shape $\overline{\mathbf{u}}$ and the eigenvector matrix $\mathbf{\Phi}$:

$$\mathbf{b}_i = \mathbf{\Phi}^T(\mathbf{u}_i - \overline{\mathbf{u}}), \qquad \mathbf{u}_i = \overline{\mathbf{u}} + \mathbf{\Phi}\mathbf{b}_i \tag{3.1}$$

To decrease the dimensionality of the representation, it is possible to use only the eigenvectors corresponding to the largest eigenvalues. In that case (3.1) becomes an approximation, with an error depending on the magnitude of the excluded eigenvalues. Furthermore, under the assumption of Gaussianity, each component of the $\mathbf{b}_i$ vectors is constrained to ensure that only *valid shapes* are represented:

$$|b_i^m| \leq \beta\sqrt{\lambda_m} \qquad 1 \leq i \leq N_I, \quad 1 \leq m \leq M \tag{3.2}$$

where β is a regularization constant, usually set between 1 and 3, according to the degree of flexibility desired in the shape model, M is the number of retained eigenvectors and λ_m are the eigenvalues of the covariance matrix.

The intensity model is constructed by computing second order statistics for the normalized image gradients, sampled at each side of the landmarks, perpendicularly to the shape's contour, hereinafter, the profile. In other words the profile is a fixed-size vector of values (in this case pixel intensity values) sampled along the perpendicular to the contour such that the contour passes right through the middle of the perpendicular. The matching procedure is an alternation of image driven landmark displacements and statistical shape constraining as in (3.2) based on the PDM, usually performed in a multiresolution fashion in order to enhance the capture range of the algorithm. The landmark displacements are individually determined using the intensity model, by minimizing the Mahalanobis distance between the candidate gradient and the model's mean.

3.2.2 Optimal features ASM

As an alternative to the construction of normalized gradients and the use of the *Mahalanobis* distance as a cost function, van Ginneken et al. [191] proposed a non-linear intensity model constructed from local image descriptors. Each pixel on the image was assigned a set of such descriptors (or *features*). Then, during matching, the landmark points are displaced along the perpendiculars to the current shape estimates to find the contours of the object of interest. However, the best displacement now will be the one for which everything on one side of the profile is classified as being outside the object, and everything on the other side, as inside of it. Optimal Features ASMs (OF-ASMs) use image derivatives as local image descriptors. The idea behind the choice of such descriptors is the fact that a function can be locally approximated by its Taylor series expansion provided that the derivatives at the point of expansion can be computed up to a sufficient order. The number of image features is reduced by sequential feature selection [100] and interpreted by a kNN classifier with weighted voting [9], to cope with the non-linearity of the texture.

3.3 Invariant optimal features ASM

This work concentrates on a generalization of OF-ASM called IOF-ASM. The modifications that we introduce in our method can be summarized as follows:

- The structure of the sampling points, as well as their interpretation, has been completely revisited (Section 3.3.3). As it will be explained in Section 3.3.4 this provides robustness with respect to shape complexities of the face.

- The local texture descriptors are replaced by irreducible Cartesian differential invariants, making the intensity model invariant to rigid transformations (Section 3.3.1).
- The kNN classifiers are replaced by multi-valued neurons (MVN) [5, 6]. The MVN is a very fast classifier whose speed is independent of the number of training samples, as opposed to kNNs (Section 3.3.2).
- During the construction of the model, different feature selection strategies can be used to tune the model towards segmentation accuracy or speed, or a compromise thereof (Section 3.3.6).

3.3.1 Irreducible cartesian differential invariants

A limitation of using the derivatives in a Cartesian framework, as features in the OF-ASM approach, is the lack of invariance with respect to translation and rotation (rigid transformations). Consequently, these operators can only cope with textured boundaries of the same orientations as those seen in the training set. To overcome this issue we introduce a multi-scale feature vector that is invariant under 2D rigid transformations.

Cartesian differential invariants describe the differential structure of an image independently of the chosen Cartesian coordinate system [67, 167, 195]. The term irreducible is used to indicate that any other algebraic invariant can be reduced to a linear combination of elements of this minimal set. Table 3.1 shows the (linear) Cartesian invariants up to second order. Notice that, for the tensor formulation, the Einstein convention is used. Hence, when an index variable appears twice in a single term, a summation over all its possible values takes place. In our context, the index variables are i and j, which can take values x or y each, the latter corresponding to the image axes.

The use of these invariants as the basis for texture description makes IOF-ASM invariant to rigid transformations. In this work we will use first and second order linear invariants at three different scales, σ = 1, 2 and 4 pixels. The zero order invariants (which correspond to the raw images seen at different Gaussian blurred scales) were not used since the differential images are expected to provide a more accurate and stable information about facial contours (edges). For example, the zero order invariant could make the texture model dependant on undesirable features, such as the color of the background surrounding the face.

3.3.2 Multi-valued neural network

In our approach we used a non linear classifier in order to label image points near a boundary or contour. Among the many available options, we have chosen the Multivalued Neurons (MVNs) mainly based on the need to improve segmentation

Table 3.1: Tensor and Cartesian formulation of invariants

Tensor Formulation	2D Cartesian Formulation
L	L
L_{ii}	$L_{xx} + L_{yy}$
$L_i L_i$	$L_x^2 + L_y^2$
$L_i L_{ij} L_j$	$L_x^2 L_{xx} + 2L_{xy} L_x L_y + L_y^2 L_{yy}$
$L_{ij} L_{ji}$	$L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2$

speed. These are very fast classifiers, since their decision is based only on a vector multiplication in the complex domain. Furthermore, a single neuron is enough to deal with non-linear problems [5, 6], which avoids the need for carefully tuning the number of layers (and neurons in each of them) that characterizes multi-layer perceptron networks.

In our approach, a MVN is assigned to each landmark, with as many inputs as the number of features selected for that landmark. All of the inputs are mapped to a complex number on the unit circle, and the argument of their weighted sum is the activation function, after appropriate scaling in order to deal with different magnitudes of the features. The number of output sectors will depend on the chosen profile size (see next Section). The reader is referred to [5] for further details.

3.3.3 Construction of the intensity model

The intensity model of IOF-ASM is based on image invariants. During the construction of the model, every image of the training set is processed in order to generate a set of *invariant images*. By an invariant image we mean that one of the invariants in Table 3.1 is computed at a specific scale for the whole image. Using these precomputed invariant images an individual intensity model is constructed for every landmark and resolution level.

The construction of the intensity model for each landmark starts by placing a rectangular grid centered at the landmark position. All invariant images are sampled at the positions defined by this grid, generated by displacing a smaller grid (sub-grid) a predefined number of positions towards each side of the landmark. Fig. 3.1 illustrates the concept: the $X_g \times Y_g$ (3×3 in the plot) sub-grid can take 5 positions, since we allow its center to depart from the landmark up to 2 pixels on each side, along the normal to the object's contour. We call the positions taken by the centers of the sub-grids as *main grid*, of size $X_G \times Y_G$ (5×1 in the drawn example). Notice that the total sampling region covered by the subgrids is $(X_G + X_g - 1) \times Y_g$ (resulting 7×3 pixels in Fig. 1). The sampled values are normalized to zero mean and unit variance across the whole sampling region to reduce the influence of global

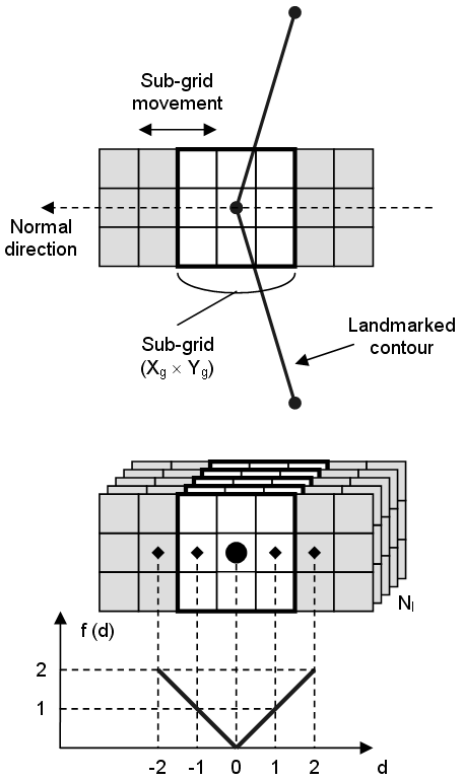

Figure 3.1: Example of sampling with five 3×3 sub-grids (top) and the corresponding labelling (bottom), based on the displacement (d) of their center from the landmark. Notice that the labelling function is $f(d) = |d|$.

illumination.

So far, for a number of positions in the neighborhood of each landmark we have the pixel intensities sampled by the sub-grid across all the (preprocessed) invariant images. That is, the extracted data for each position of the main grid can be thought as a cuboid of size $X_g \times Y_g \times N_{InvAll}$, where X_g, Y_g are the dimensions of the sub-grid and N_{InvAll} is the number of invariant images. Each of the cuboids is labelled according to the distance from its center to the landmark, as shown on the bottom part of Fig. 3.1. The typical plot of the labels as a function of the sub-grids displacements will take a shape of letter "V". Its vertex will be located at the landmark position.

After labelling all the training images, the labelled cuboids are used to train their corresponding MVN *texture classifier*. Note that for each landmark there is an

independent MVN with $\frac{1+X_G}{2}$ output sectors[1]. When used for classification, the MVN will return the distance to the most likely position for the landmark (according to its training) in the (continuous) interval defined by $[-\frac{1}{2};\ 1+\frac{X_G}{2})$ (notice that the interval of interest is $[0;\frac{1+X_G}{2}]$ and there is a $\frac{1}{2}$ extension to each side because integer-valued labels are mapped to the centers of the MVN sectors [5]).

3.3.4 Shape complexities

Let us revisit for a moment the OF-ASM. Its training is based on a landmarked set of images for which all derivative images were computed and described by local (histogram) statistics. Once the texture classifiers are trained, they would be able to classify a point as inside or outside the region of interest based on the texture descriptors (the features). Therefore, labelling inside pixels with 1 and outside pixels with 0 and plotting the labels corresponding to the profile pixels, the classical step function is obtained, and the transition will correspond to the landmark position.

Nevertheless, there are a couple of reasons why this may not happen. The first one is that certain regions of the object can be thinner than the size of the grid, and then the correct labelling of the points would look more like a bar rather than like a step function. An illustrative example arises when the square grid is placed over the eye or eyebrow contours (Fig. 3.2). Moreover, in a multiresolution framework, image sub-sampling contributes to “step over" these structures. Another problem is that the classifiers will not make a perfect decision, so the labelling will look much noisier than the ideal step or bar. Additionally, Fig. 3.2 illustrates how, for certain landmarks where there is a high contour curvature (i.e mouth/eyes/eyebrows corners), most of the grid points would lie outside the contour, promoting quite an unbalanced training of the classifiers.

The IOF-ASM has been designed to deal with these problems: once the learning process is completed, the MVNs should be able to tell the distance of a given cuboid with respect to the correct landmark position. Therefore, the typical plot of the profiles will be a “V", with its minimum (the vertex) located at the landmark position, irrespective of which region is sampled and its width relative to the grid size.

3.3.5 Model matching

During matching, the best fit to the “V" shape is searched for at every landmark. The subgrids are moved over a more extense area than during training, to allow for several *candidate profiles* to select the best V from. An example of this is shown in Fig. 3.3, where the main grid is allowed to move three pixels to each side of the

[1]Since the sub-grids are made symmetric with respect to their center, X_G is an odd integer and, therefore, the resulting number of output sectors is also integer.

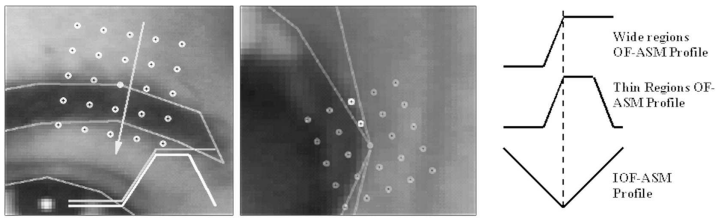

Figure 3.2: A typical eyebrow image and a 5x5 grid with the arrow indicating the normal to the contour (left); The same grid over the mouth corner, where only 3 points lie inside the lip (center); and the plots of the typical profiles for OF- and IOF-ASM (right) for the eyebrow image showed on the left.

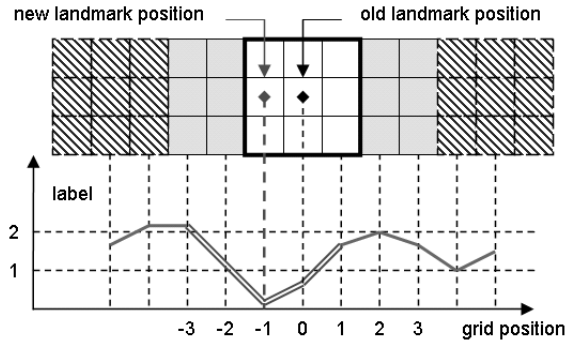

Figure 3.3: Looking for new landmark positions during model matching.. A number of sub-grids is sampled over the image and classified by the MVN. The bottom plot shows the labels assigned to the sub-grids for each position. The landmark point will be displaced to the position that best fits to the V-shape.

current landmark position. Thus, there are seven main grids in the plot, and each of them contains five sub-grids (so the profile to search for is a five-pixel wide V). The outputs of the texture classifier (the labels for each sub-grid) show that the group of points that best fits the V is centered at -1 (indicated with a double line in Fig. 3.3), one pixel away from the current landmark position. That would be the updated position for that landmark.

The fact that all sub-grids are labelled with the distance to the landmark allows for introducing a robust estimation metric [84]. The best position for the landmark is now the one which minimizes the profile distance to the ideal V, excluding the *outliers*. In this context, an outlier is a point on the profile whose distance to corresponding point on the ideal V is greater than one. The later can be easily understood by noticing that such a point is suggesting a different position to place the landmark (i.e. its distance would be smaller if the V is adequately displaced). If the

number of outliers exceeds 1/3 of the profile size, then the image model is regarded as not trustworthy and the distance for that position is set to infinity. Otherwise, preference is given to the profiles with fewer outliers. The objective of matching the intensity model is to find such a k that minimizes

$$f(k) = N_{OL}(k) + \frac{\sum_{i=1}^{N_P - N_{OL}(k)} |p_i(k) - v_i|}{N_P - N_{OL}(k)} \tag{3.3}$$

where k are the different candidate positions for the landmark, $N_{OL}(k)$ is the number of outliers, N_P the profile size, and p_i and v_i are the *i-th* components of the *k-th* input profile, $\mathbf{p}(k)$, and ideal (V) profile, respectively.

3.3.6 Feature selection

The computational load of the segmentation process can be associated to two main tasks: the computation of invariant images and the iterations of the matching procedure. The invariants computation time is proportional to the image size: its complexity is roughly $O(N_{InvAll} \times I_H \times I_W)$, where I_H and I_W are the height and width of the image, respectively. On the other hand, the iterative process complexity is proportional to the number of landmarks, the number of features and the number of iterations. Notice that the (total) number of features is the size of the sub-grids multiplied by the number of invariants. Thus, it is clear that selecting a subset of the available features will speed-up the segmentation. Additionally, if this selection determines that some invariant images will not be used at all by any landmark, then there will be a further speed-up by skipping their precomputation.

Among feature selection methods, wrapper greedy search seems to be particularly computationally advantageous and robust to over-fitting [76]. In IOF-ASM we use a Sequential Backwards Selection (SBS) [100] without retraining the classifier while evaluating the features to exclude. This is combined with a voting strategy to determine the exclusion of the same invariants for all landmarks at once, thus avoiding their calculation at the preprocessing stage of matching. The algorithm, which is executed independently for each resolution level, is outlined below. The iterations start with all of the available invariants, N_{InvAll}, excluding one invariant at each step until it reaches N_{InvF}, the desired (*final*) number of invariants:

Let us concentrate on lines 6 to 10. The learning process for a MVN can involve some thousands of passes through all training samples while the classification score is computed in a single pass, so its computational load can be considered negligible. Then, if the texture classifiers are not re-trained while deciding which invariant to discard, the solution is sub-optimal but much faster. While the feature selection does not constitute a step of the matching algorithm, it is of practical necessity to speed-up this process. The improvement due to the absence of retraining at line 8

Algorithm 1 Joint feature selection with no retraining

1: $n = N_{InvAll}$
2: **while** $n > N_{InvF}$ **do**
3: Initialize voting array to zero
4: **for** $l = 1$ to number_of_landmarks **do**
5: Train texture classifier (with n invariants)
6: Compute classifier score
7: **for** $i = 1$ to n **do**
8: Compute score without *i-th* invariant
9: **end for**
10: Update voting array based on saved scores
11: **end for**
12: Eliminate most voted invariant
13: $n = n - 1$
14: **end while**

is:

$$\eta \simeq \frac{SBS_{NR} \quad speed}{SBS \quad speed} = \sum_{n=N_{InvF}}^{N_{InvAll}} n - 1 \tag{3.4}$$

where SBS_{NR} stands for SBS with no retraining. That is, at each elimination step there is only one re-training of the classifier (with the current set of features, at line 5) instead of one re-training per feature[2]. It can be seen that the speed improvement grows easily by 1 or 2 orders of magnitude even for small sets of invariants.

The other key point of the algorithm is at line 10. The invariant image to be discarded at each step will be the same for all landmarks, selected according to a weighted voting strategy [7]. Let $s_{0,l}$ be the initial classifier score for landmark l (computed at line 6) and $s_{k,l}$ the same score after excluding the k-th invariant (line 8). The classifier scores range from 0 (complete failure) to 1 (perfect success). Then,

$$\Delta_{k,l} = s_{0,l} - s_{k,l} \tag{3.5}$$

measures how much does the k-th invariant affect texture classification for landmark l. Actually, we define the index k such that the invariants are sorted in ascending order of $\Delta_{k,l}$. Then, every landmark l assigns $v_{k,l}$ votes to each invariant k according to:

$$v_{k,l} = (1 - \Delta_{k,l})\ 2^{-k} \tag{3.6}$$

It can be seen that the lower $\Delta_{k,l}$, the less important is the *k-th* invariant for the *l-th* landmark and, therefore, the more votes are assigned to the exclusion of that invariant. The negative exponential balances the voting privileges among all landmarks

[2]Under the above considerations, each elimination cycle skips $n - 1$ re-trainings of the classifier. Since the time for computing the classification score is negligible, this leads immediately to (3.4).

(i.e. each landmark will at most influence just a few invariants, regardless of the values of $\Delta_{k,l}$). The invariant eliminated at each step is then the most voted one:

$$\underset{k}{\operatorname{argmax}} \sum_{l} (1 - \Delta_{k,l})\ 2^{-k} \tag{3.7}$$

3.4 Experimental evaluation

The performance of the proposed method was compared to that of the ASM and OF-ASM schemes. Datasets from three different databases were used, namely the AR database [126], XM2VTS database [133], and Equinox database (visible band images from [171]).

Table 3.2: Composition of the employed datasets and the different groups into which they were divided.

Database	AR	Equinox	XM2VTS
Total identities	133	91	295
Images per person	4	6	8
Total images	532	546	2360
Landmarks per image	98	98	64
Users	90	62	200
· training images	180	186	800
· test images	90	124	400
· eval. images	90	62	400
Test Impostors	32	21	70
· images	128	126	560
Eval Impostors	11	8	25
· images	44	48	200

The performance was tested in terms of segmentation accuracy and identity verification scores. The Configuration II of the Lausanne protocol [133] was used for the XM2VTS database, while AR and Equinox datasets were divided accordingly to make verification scores comparable. The individuals in each group were randomly chosen, making sure to have the same proportion of facial expression in all of them. Table 3.2 summarizes the resulting number of images on each group as well as the templates used to landmark each dataset. The AR and Equinox datasets have been landmarked with our own 98-points template [177], while XM2VTS landmarks were obtained from [1] with a 68-points template

The Equinox images were enlarged by a factor of 2.2 : 1 such that the average distance between the centers of the eyes matched that of AR dataset (approx. 115

Table 3.3: Parameters used to build the statistical models

Parameter	ASM	OF-ASM	IOF-ASM
Main grid size	n/a	5×5	$X_G = 7, Y_G = 1$
Sub-grids size	n/a	$\alpha = 2\sigma$	$X_g = 7, Y_g = 5$
Profile length (p)	17	9	7
Resolutions	4	5	5
Iterations	12	12	12
Image properties	n/a	L, L_x, L_y	$L_{ii}, L_i L_i$
		L_{xx}, L_{xy}, L_{yy}	$L_i L_{ij} L_j, L_{ij} L_{ji}$
Blurring scales (pix)	n/a	$\sigma = 1, 2, 4$	$\sigma = 1, 2, 4$

pixels). The XM2VTS images were kept unchanged since the needed resizing factor was less than $1.13 : 1$. In all of the experiments that will be presented, only luminance information has been used.

The following sections evaluate the algorithm in terms of segmentation accuracy, rotation invariance and identity verification, as well as the influence of the joint features selection strategy explained in Section 3.3.6.

3.4.1 Segmentation accuracy

We constructed an ASM, an OF-ASM[3] and an IOF-ASM model for each of the three datasets, and tested their performance on all datasets. The models were built always from the training images of the *users* group (see Table 3.2), such that they can also be used in identity verification tests. In all cases, the image model was allowed to search within ± 3 pixels along the profiles (on each iteration) and β was set to 1.5 (see (3.2)). The rest of the parameters are shown in Table 3.3. They were chosen to obtain the best ASM results over the AR dataset and were kept unchanged for the other databases and algorithms, whenever possible. For specific OF-ASM and IOF-ASM parameters, segmentation speed was given more importance than accuracy. It should be noted that, due to the different structure of the sampling grids in OF-ASM and IOF-ASM, the parameters of Table 3.3 make their sampling regions (per landmark) coincident for the smallest scale of OF-ASM[4].

The face location was assumed to be roughly known from a (previous) detection

[3]For OF-ASM the profiles to search for were modified according to the training set statistics, since the ones proposed originally in [191] performed too poorly to allow for comparison of results (see Section 3.3.4).

[4]The sampling region of OF-ASM is $(p + 2\alpha) \times (2\alpha + 1)$, so its minimum size is 13×5, when $\sigma = 1$. The sampling region of IOF-ASM is $(X_G + X_g - 1) \times Y_g$, that is 13×5 pixels independently of the blurring scale.

step[5] and all available features were used for the segmentation. Feature selection will be covered in Section 3.4.4.

Table 3.4 shows the point-to-curve segmentation error, based on the distance from the landmarks obtained by the segmentation, measured perpendicularly to the curve defined by the manual annotations. The displayed values correspond to the average over all landmarks and all segmented images, with their corresponding standard error. The errors for each face were normalized dividing by 0.01 of the distance between the centers of the eyes (from the manual annotations) to make segmentation error comparable among faces of different sizes.

Table 3.4: Point-to-curve segmentation error, normalized to the distance between eye-centers.

Training with	Model	Segmenting AR	Segmenting Equinox	Segmenting XM2VTS
AR	ASM	2.42 ± 0.06	3.74 ± 0.07	9.62 ± 0.20
	OF-ASM	2.55 ± 0.12 **(+5.4 %)**	12.2 ± 0.41 **(+227 %)**	6.04 ± 0.06 **(-37.2 %)**
	IOF-ASM	1.63 ± 0.03 **(-33.2 %)**	3.59 ± 0.07 **(-4.2 %)**	5.22 ± 0.10 **(-45.8 %)**
Equinox	ASM	4.72 ± 0.09	2.56 ± 0.05	13.9 ± 0.27
	OF-ASM	7.24 ± 0.21 **(+53.4 %)**	2.17 ± 0.07 **(-15.2 %)**	11.3 ± 0.09 **(-18,7 %)**
	IOF-ASM	4.16 ± 0.10 **(-11.8 %)**	1.92 ± 0.03 **(-25.2 %)**	8.35 ± 0.17 **(-39.7 %)**
XM2VTS	ASM	5.17 ± 0.14	4.45 ± 0.10	3.07 ± 0.08
	OF-ASM	7.92 ± 0.28 **(+53.2 %)**	13.9 ± 0.34 **(+169 %)**	2.13 ± 0.03 **(-30.6 %)**
	IOF-ASM	4.21 ± 0.12 **(-18.6 %)**	4.06 ± 0.11 **(-8.8 %)**	2.03 ± 0.02 **(-33.8 %)**

Below the segmentation errors for OF-ASM and IOF-ASM, we show the difference with respect to the ASM results as a percentage. It can be seen that IOF-ASM performed the best in all cases. We also observe a high degree of consistency in the results over the diagonal of the table, that is, when the segmented images belong to the same database used to construct the models. In these cases, the IOF-ASM

[5]The model is always initialized at around 90% size of the average face in the corresponding database. In this way, the initialization usually falls inside the face region, then reducing the background effects which regard mostly to the detection step. The average point-to-point error of the initialization was of 13.4 and 17.4 pixels for the AR and Equinox databases, respectively, while the greater size variations of the XM2VTS database made this initialization error grow up to 47 pixels (on average) in this database.

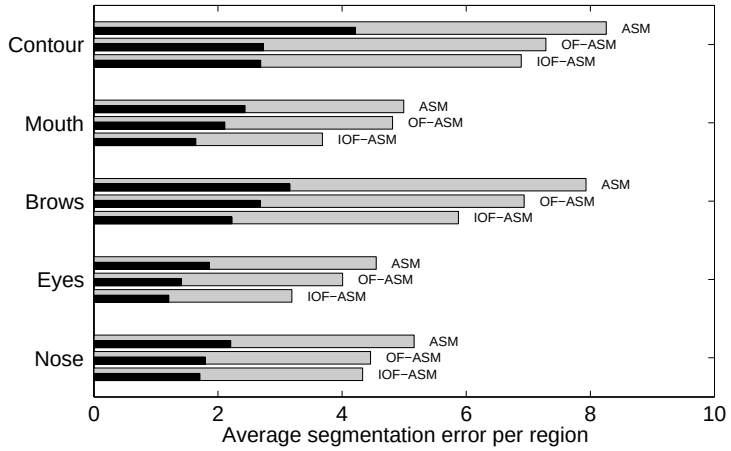

Figure 3.4: ASM, OF-ASM and IOF-ASM point-to-point (light) and point-to-curve (dark) segmentation errors per region. The values are in pixels, normalized with respect to the inter-eye distances and averaged over AR, Equinox and XM2VTS datasets together, each segmented with their own models.

approach always outperformed ASM by about 30%. In Fig. 3.4 we show the average segmentation errors over the table's diagonal, divided into the different facial regions. It can be seen that IOF-ASM always performed better, with both point-to-point and point-to-curve error metrics.

Fig. 3.5 shows further comparison of ASM and IOF-ASM accuracy when varying the PDM regularization constant β. It can be seen that, as β increases, the difference between the error of both models tends to grow. At the same time, the PCA reconstruction error introduced by the PDM decreases, which means the segmentation relies more on the image model precision. This behavior enforces the hypothesis of performance improvement in favor of IOF-ASM.

The value of the regularization constant greatly influences the segmentation performance. As shown in the curves displayed for ASM, if the PDM is given too much freedom the segmentation error could exceed even the error obtained at $\beta = 0$; when only a similarity transformation of a fixed shape (the mean) is allowed. Examples of segmentation results are shown in Fig. 3.6. When β is increased to 3 the shape is clearly less restricted by the PDM, and the result achieved by the ASM is not a plausible face.

Away from the diagonal of Table 3.4, when different databases were used for training and testing, the behaviors were more random, but IOF-ASM was still the best, achieving statistically significant improvements with respect to both ASM and

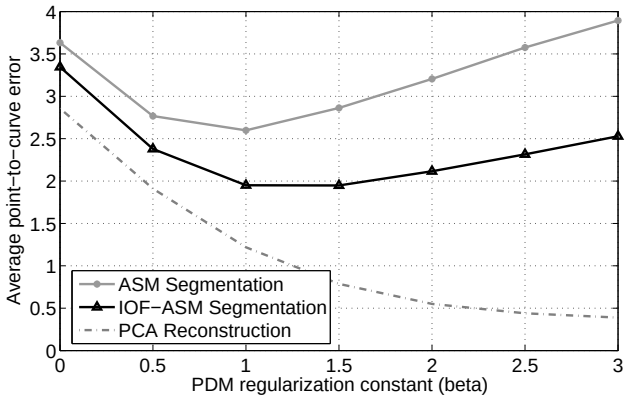

Figure 3.5: ASM and IOF-ASM point-to-curve segmentation errors while varying the regularization constnat of the PDM. The IOF-ASM improvement progressively grows from 8% at $\beta = 0$ to more than 35% at $\beta = 3$. The reconstruction error due to the regularization constraints of the PDM is also shown.

OF-ASM in all cases[6]. An underlying conclusion from these experiments, however, is that none of the models was able to preserve the same accuracy when running the segmentation on a database different to the one it has been trained with.

3.4.2 Invariance to image rotations

It was emphasized in Section 3.3.1 that the IOF-ASM features extracted from the images are invariant to rigid transformations. ASM exhibits the same invariance, but OF-ASM does not. To verify this fact we performed the segmentation experiments at the finest resolution on the AR dataset by rotating the images from -150 to +150 degrees. The PDM was constructed from the rotated images, such that the starting shape (based on the *mean shape*) was also rotated. However, the image models were not changed (i.e. they were based on the non-rotated images) so that their invariance was the only thing to be tested.

The results of the experiment are presented in Fig. 3.7. For each method, all segmentation errors were divided by the value obtained when segmenting the non-rotated images. Therefore, the three methods were scaled differently according to their respective accuracy, and the plot demonstrates only the relative influence of the rotation angle on each of them.

As expected, there is a clear increase of the segmentation error in the OF-ASM as the rotation angle departs from zero. On the other hand, the ASM and IOF-

[6]The t-test showed confidence larger than 99% in all cases except when comparing IOF-ASM and ASM on the Equinox database training with the AR database, where the confidence was around 97%

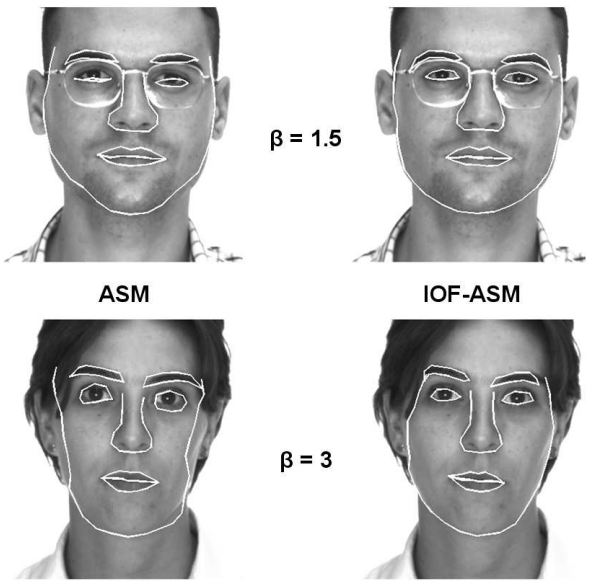

Figure 3.6: Typical segmentation results of ASM and IOF-ASM on images from the AR database for different values of the PDM regularization constnat (see (3.2)).

ASM performances are only affected by the numerical approximations due to the discrete nature of the image. That is, the samples for the rotated versions of the image were computed by interpolation, except for ± 90 degrees, where the error attains again its minimum value. The lack of invariance can be substantial due to numerical approximations when doing multi-resolution, since the chosen low-pass filters usually depend on the orientation of the axis. Care must be taken both when implementing and testing invariant methods in that context.

3.4.3 Sub-grid size

The size of the sub-grids is an important parameter of IOF-ASM. It can affect both the accuracy and the speed of the segmentation process, as demonstrated in Fig. 3.8. As explained in Section 3.3.3, X_g and Y_g can take only odd values. Hence, we repeated the segmentation test on the AR database for all the possible sub-grid sizes from 3 to 9 pixels in both dimensions (i.e. 3×3, 3×5, etc).

The first conclusion from these experiments was that accuracy and speed changed as a function of the total number of points of the sub-grids, being almost insensitive to the swap between X_g and Y_g. For this reason, the horizontal axis of Fig. 3.8 is labelled with the product $X_g \times Y_g$.

Analyzing the segmentation error, it can be seen that very small sub-grid sizes

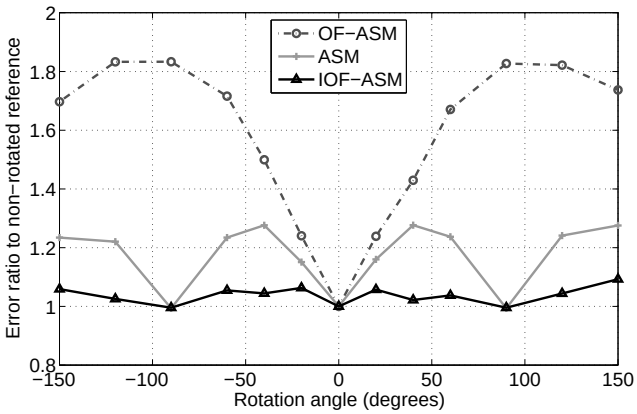

Figure 3.7: ASM, OF-ASM and IOF-ASM point-to-curve segmentation error for different rotation angles on the AR database. The models were constructed with the non-rotated images, whose segmentation error was taken as reference. The plot shows the ratio of the segmentation errors (for the different rotation angles) to the (non-rotated) reference.

degrade the accuracy. However, for sub-grids of 25 or more points, the differences are very subtle. Statistical significance was observed only for the sub-grids smaller than 25 points.

Regarding the segmentation time, its variation was linear with respect to the number of pixels[7]. The strong linearity of the curve allows to clearly distinguish between the time consumed by the iterative fitting and by the preprocessing (the extrapolation of the line to intersect the vertical axis for a hypothetical zero-size sub-grid). This preprocessing time, of approximately 3.5 seconds, is mainly due to the computation of the invariants.

The size of 7×5 points chosen for the experiments in Table 3.3, is within the flat region of accuracy, and allows for a direct comparison with OF-ASM, since the total region analyzed for each landmark will match the one of OF-ASM for $\sigma = 1$ (see Section 3.4.1). However, the displayed curves suggest that smaller sub-grid sizes could accelerate the segmentation up to 15% without compromising the accuracy.

3.4.4 Feature selection

The price paid for the accuracy improvement of IOF-ASM reported in Section 3.4.1 is a decrease in segmentation speed due to the more complex image processing

[7]The reported segmentation times were measured on a non optimized implementation of the algorithms in C++, on an AMD Athlon running at 2 GHz. Significantly better times should be possible, especially if optimizing the calculation of invariants.

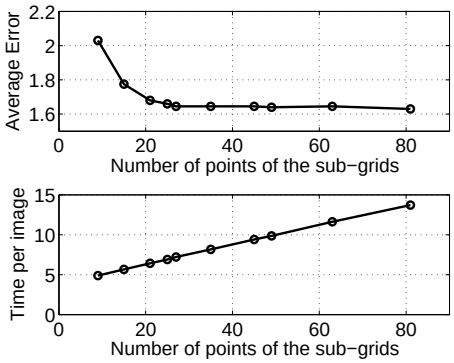

Figure 3.8: Normalized point-to-curve segmentation error (top) and segmentation time, in seconds (bottom), while varying the number of points of the sub-grids. The displayed curves show the average over the 532 images from the AR dataset.

compared to ASM. However, the method can be accelerated by reducing the number of features.

Using the training images (from the *users* group) of AR database, we performed the feature selection as explained in Section 3.3.6, and reduced the number of invariant images from 12 to 4. We constructed several models with different number of invariants and compiled the segmentation results in Table 3.5. From left to right, the columns show the number of invariants used to build the model, the point-to-curve errors averaged over the AR database, over all images from the three available datasets (a total of 3438), and the segmentation time per image. The ASM and OF-ASM results are also shown and the percentages are computed by taking the IOF-ASM with all the features as a baseline.

It can be seen that the smaller the number of invariants (N_{InvF}) we use to build the model, the higher the segmentation error and the lower the segmentation time. This behavior was very clear in segmenting the AR database, since part of its images were used to guide the feature selection process. When moving to other datasets, where the best set of features may in general be different, there was some unexpected decrease of the average error with six invariants, but the tendency was clearly the same.

Therefore, N_{InvF} can be chosen to make the model faster or more accurate. Furthermore, the influence of excluding invariants from the model can be evaluated prior to the segmentation experiments, during training. The continuous lines in Fig. 3.9 show the classifier scores averaged over all the landmarks for each resolution level. Notice the clear correlation between them and the pairs of bars showing the segmentation accuracy (left) and the segmentation time (right).

The results displayed suggest that all of the used invariant images brought valu-

Table 3.5: Segmentation accuracy and speed for different number of invariants, over 3438 images from AR, Equinox and XM2VTS datasets

Model	N_{InvF}	Avg error AR	Avg error ALL	Time per image
ASM	n/a	2.42 ± 0.06	7.58 ± 0.16	0.70 s
OF-ASM	n/a	2.55 ± 0.12	6.47 ± 0.13	> 100 s
IOF-ASM	12	1.63 ± 0.03	4.40 ± 0.09	8.13 s
	10	1.66 ± 0.03 **(+1.8 %)**	4.43 ± 0.09 **(+0.7 %)**	7.31 s **(-10 %)**
	8	1.71 ± 0.03 **(+4.9 %)**	4.59 ± 0.09 **(+4.3 %)**	6.39 s **(-21 %)**
	6	1.82 ± 0.04 **(+11.7 %)**	4.49 ± 0.09 **(+2.1 %)**	5.62 s **(-31 %)**
	4	1.87 ± 0.04 **(+14.7 %)**	5.79 ± 0.11 **(+31.6 %)**	4.09 s **(-50 %)**

able information to the texture classifiers. The most discriminant feature in our experiments was $\{L_{ij}L_{ji}, \sigma = 1\}$, which was among the 4 non-excluded features in all the five resolution levels.

3.4.5 Step by step analysis

It is interesting to go back to the hypothesis made at the beginning if Section 3.3. It was stated that IOF-ASM would provide a number of improvements with respect to OF-ASM, which were attributed to specific components (or changes) of the new method. To verify this, we constructed intermediate versions between OF-ASM and IOF-ASM, which are summarized in Table 3.6. Three measures were computed to compare their performance, by using the AR dataset:

- e: Point-to-curve segmentation error, averaged over all landmarks and all images, with their corresponding standard error. The error values are normalized as in Table 3.4.
- r: Rotation variance ratio. The experiments of Section 3.4.2 were repeated for the original images (no rotation) and rotated versions at ± 60 and ± 120 degrees. Then, r was computed as the ratio from the average segmentation error of the rotated images to the error of for the non-rotated images. Hence, as in Fig. 3.7, if a model is rotation-invariant, then r approaches 1.
- t: Segmentation time per image (on average), under the same conditions of Sections 3.4.3 and 3.4.4.

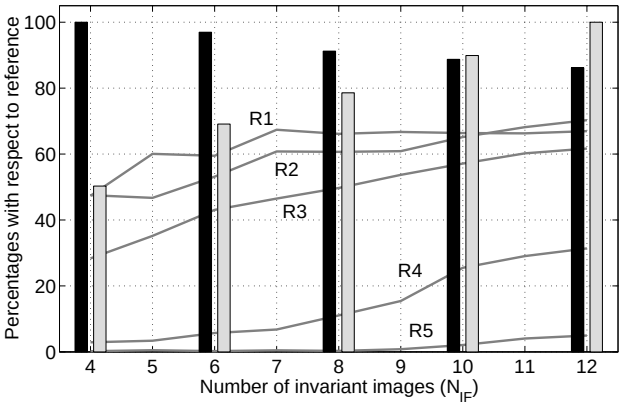

Figure 3.9: Feature selection statistics from the AR model reducing N_{InvF} from 12 to 4. The continuous lines show the texture classifiers scores after training, averaged over all landmarks at each resolution level. The bars display the segmentation accuracy (black) and the segmentation time (gray) for the AR dataset (see Table 3.5). The maximum values of the bars are used as reference (fixed to 100%) and the rest are rescaled accordingly.

On each row the change with respect to the previous one is highlighted in bold letters, as well as the most affected measure. The results are very consistent with the expectations drawn in Section 3.3. The exchange of the kNN classifier by the MVN (second row) reduced the segmentation time in one order of magnitude, with a small accuracy loss (no statistically significant in this case). The use of invariants instead of derivatives (third row) did not affect the segmentation error, but clearly achieved rotation invariance.

The remaining change at this step is the replacement of the OF-ASM profiles by the V-shape of IOF-ASM, including the modification of the grids structure (see Section 3.3.4). The resulting model is the IOF-ASM proposed in this paper (row five). However, an intermediate step is shown in row four, which does not use the outliers concept of (3.3) but a simple absolute distance. It can be seen that both steps considerably reduced the segmentation error (the results are statistically significant with confidence greater than 99%).

3.4.6 Identity verification

Having demonstrated that IOF-ASM is more accurate in segmenting facial images, there remains the question of whether or not it will improve recognition as well. We tested verification scores using both shape and texture classifiers in the three datasets with the segmentation performed by their corresponding model. The de-

Table 3.6: Intermediate steps from OF- to IOF-ASM

Model	Classif.	Features	Profiles	Results
OF-ASM	kNN	Derivatives	OF-ASM	$e = 2.55 \pm 0.12$ $r = 1.76 \pm 0.03$ $t > 100s$
n/a	**MVN**	Derivatives	OF-ASM	$e = 2.72 \pm 0.11$ $r = 1.69 \pm 0.04$ $\mathbf{t = 9.78s}$
n/a	MVN	**Invariants**	OF-ASM	$e = 2.68 \pm 0.10$ $\mathbf{r = 1.03 \pm 0.01}$ $t = 8.54s$
n/a	MVN	Invariants	**V-shape**	$\mathbf{e = 1.83 \pm 0.04}$ $r = 1.06 \pm 0.01$ $t = 8.15s$
IOF-ASM	MVN	Invariants	**Robust-V**	$\mathbf{e = 1.63 \pm 0.03}$ $r = 1.05 \pm 0.02$ $t = 8.13s$

velopment of a state-of-the-art classifier was beyond the scope of this paper. In both cases we applied Z-normalization [110] and then used a whitened correlation classifier, known to be a good choice for PCA-based metrics [145].

In Table 3.7, the landmark points found during the model matching were used to construct a mesh by means of a Delaunay triangulation of the whole face. This mesh permits to establish a mapping for the texture of the face from this specific shape to the mean shape of the training set. Using this mapping the texture was warped onto the mean shape and sampled into a texture vector [106]. In order to parameterize the texture, a model was constructed by applying PCA to all the vectors from the training images of the database. Then the biometric parameters were obtained by projecting the texture vector onto the subspace of this model.

The results in Table 3.8 were obtained using the PCA parameters of the shape model. Only the most significant modes were taken into account, such that 95% of the total variance was explained. The error rates were computed according to the protocols described in Table 3.2, using the *evaluation* sets Equal Error Rate (EER) [21] to fix the working point of the classifier and compute the False Acceptance (FAR) and False Rejection (FRR) rates from the *test* sets. The tables show the average of both metrics (HTER, for Half Total Error Rate) plus a 90% confidence interval, computed using the *test of two proportions* as in [12].

The results show that IOF-ASM outperforms the other methods in all cases, except when using shape parameters on the Equinox dataset. In that case, a lower

Table 3.7: Identity verification scores using texture parameters

Database	Metric	ASM	OF-ASM	IOF-ASM
AR	EER (Eval)	3.8%	5.6%	2.8%
	HTER (Test)	4.5% ±1.9%	5.8% ±1.8%	4.2% ±1.8%
Equinox	EER (Eval)	0.1%	0.6%	$< 0.1\%$
	HTER (Test)	1.2% ±1.6%	1.6% ±1.6%	0.5% ±0.9%
XM2VTS	EER (Eval)	2.7%	2.4%	1.0%
	HTER (Test)	2.6% ±0.8%	2.3% ±0.7%	1.3% ±0.6%

Table 3.8: Identity verification scores using shape parameters

Database	Metric	ASM	OF-ASM	IOF-ASM
AR	EER (Eval)	16.6%	20.0%	15.6%
	HTER (Test)	17.7% ±3.4%	19.5% ±3.3%	11.9% ±2.5%
Equinox	EER (Eval)	6.4%	14.5%	9.7%
	HTER (Test)	11.5% ±3.9%	16.2% ±4.1%	11.6% ±3.6%
XM2VTS	EER (Eval)	13.8%	13.7%	10.0%
	HTER (Test)	14.0% ±1.4%	13.4% ±1.4%	9.1% ±1.1%

EER is achieved by using ASM, but that behavior does not generalize to the test set (HTER), where the scores are the same as those of IOF-ASM. It must be pointed out that, in most cases, the differences in error rates are not statistically significant, due to the limited number of images available. However, the trend in the three datasets is consistent and indicates an improvement in the verification task possibly due to the more accurate segmentation. Additionally, the work by Kang et. al. [91] allows for the comparison of IOF-ASM with similar approaches. In that work, they outperform the best distance measures using the eigenfaces approach [188] obtaining a 2.6% EER on the XM2VTS database. This is similar to the ASM performance shown in Table 3.7, but clearly worse than IOF-ASM. In the same work, however, more sophisticated classification schemes demonstrate better identity verification rates.

3.5 Discussion

The results presented in the previous Section show a significant accuracy improvement of the IOF-ASM with respect to its predecessors, namely the ASM and OF-ASM. To put these results into a more general context, a wider comparison with other methods would be desirable. The task, however, is not easy. On one hand, segmentation accuracy is usually tested against manual annotations, which are subjective and not widely available. Only recently some large-size facial datasets have been annotated and made freely available[8]. This fact has led researchers to evaluate their methods on different databases, with different image sizes, annotation templates and/or number of samples, hampering consistent comparisons.

Moreover, there is no universally accepted standard for the metric to be used. Researchers have reported segmentation results as the percentage of pixels correctly identified inside a region [116], or the fraction of test images correctly detected within a threshold [58] [50], to cite some. The most popular measurement is the Euclidean distance, in its point-to-point or point-to-curve variants. Despite the former is the most intuitive, the latter usually provides greater correlation between the error value and the failure of the model when the objective is to determine the boundaries between regions. A clear example are the points on the contour of the face, where the exact location of the landmarks does not really matter. What matters is the way the curve fits into the face boundary. As opposed to that, we can also find points in the face whose exact location is important, such as the corners of the eyes or the mouth.

Considering the above restrictions, we have gathered in Table 3.9 a list of the results reported in the literature that are possible to compare with our experiments. The error values (given in pixels) were *corrected* dividing by the average distance between the eyes, obtaining an estimation of the *normalized error*. In this way we made these results comparable to the ones presented in the previous Section. The most direct comparison is probably a work of Scott et al. [169]. They tested several AAM approaches on (almost) the whole XM2VTS database, with the same kind of annotation template as the ones used here. Depending on the technique, their errors range from 3.9 to 5.4 pixels (no correction is needed here, since the XM2VTS images have an average inter-eye distance of approximately 100 pixels); significantly higher, even than our implementation of ASM. This result was somehow expected [40], although the difference should not be that high. The explanation is probably the different choice of the parameters and the more challenging initialization used in [169].

In [42] the accuracy of AAMs is tested with a better initialization, resulting from a previous estimation of the correct pose. An average point-to-point error of 4.0 pixels is obtained, on approximately 200 pixels wide faces. To determine

[8]Some annotated datasets can be found at http://www.isbe.man.ac.uk/~bim/

the correction factor we computed the ratio between the inter-eye distance and the face's width on our data sets, which was around 2.6 (on average). In this way, the normalized error would be 5.2 pixels, but this value does not include a 1.6% of the test data for which the algorithm did not converge (average error greater than 7.5 pixels). Therefore, 5.2 pixels is the (estimated) lower bound for the segmentation error.

In [200] Elastic Bunch Graph Matching (EBGM) is used to segment facial images with a less dense template that includes the whole head. An average point-to-point error of 1.6 pixels is reported on a 432 images database, in which the distance between the center of the eyes is not higher than 30 pixels. Therefore, the normalized error in this case becomes greater than 5.3 pixels, which suggests a slightly lower performance than our IOF-ASM. However, the much lower resolution of their images and the presence of slightly in-depth faces rotation hampers any direct comparison. On the other hand, their less dense template implies some benefit when using point-to-point distance, since this allows for choosing mainly clearly defined points. As opposed to that, more dense templates are meant to define curves along the boundaries of different regions. It is often the case that the intermediate points along the boundaries have a very ambiguous location (i.e. the face silhouette) and the automatic segmentation differs from the manual annotations along the boundary, wrongly increasing the computed distance.

Tamminen and Lampinen [182] segmented images from the IMM database [139], using a variant of AAM based on Gabor filters. The average inter-eye distance on this database is almost 126 pixels, so the normalized error suggests that their results are very good. Unfortunately, they report tests on only 37 images, a much smaller dataset than the other works compared here.

3.6 Summary and conclusions

In this paper a new segmentation method has been presented to solve some limitations of its predecessor, the OF-ASM approach. The main contributions introduced here are an increased segmentation accuracy, invariance to rigid transformations, ability to deal with shape complexities (such as multiple embedding) and the speed-up of the segmentation process.

The IOF-ASM was compared with its two predecessors over three different datasets (almost 3500 images). Moreover, we gathered data from different segmentation methods to put our results in a more global context. It was shown that the achieved accuracy is comparable to other state-of-the-art algorithms, and that differences become smaller when similar techniques are employed (i.e. the Gabor-based AAM in [182]). That is, by means of using more elaborated descriptions of the texture it is possible to increase the accuracy of the segmentation. In this regard, our method provides a generic framework, since it can be extended to any new set

Table 3.9: Comparison with the segmentation errors reported by other researchers

Method and Paper	Data Sets Details	Reported error (in pixels)	Normalized error (estimated)
IOF-ASM	AR-XM2-EQX 3438 images see Section 3.4	1.95 point-curve 4.82 point-point	1.95 point-curve 4.82 point-point
AAM [169]	XM2VTS 1817 images 720×576 pix	3.9 to 5.4 (point-curve)	3.9 to 5.4 (point-curve)
AAM [42]	400 images	4.0 pixels (point-point)	$>$ 5.2 pixels (point-point)
EBGM [200]	Bochum [102] 432 images 128×128 pix	1.6 pixels (point-point)	$>$ 5.3 pixels (point-point)
AAM [182]	IMM 37 images 640×480 pix	2.78 point-curve 5.57 point-point	2.21 point-curve 4.42 point-point

of local descriptors. It was shown that using just differential invariants up to the second order is enough to obtain a very good performance.

The IOF-ASM has been designed to provide segmentations by means of dense annotated templates. Thus, the different regions of an object can be identified and analyzed by further processing. We have demonstrated this by performing identity verification experiments, obtaining results comparable to the fully automatic methods reported in [131], although we used only basic techniques to classify the identities.

The price paid to increase accuracy is, in general, a higher computational load. The banks of filters applied to the image significantly slow down the matching process. To reduce this effect we have shown that feature selection can save up to 50% of computational time while degrading accuracy by only about 15%. Different trade-offs between speed and accuracy are also possible.

CHAPTER 4

Projective Active Shape Models for Pose-variant Image Analysis of Quasi-Planar Objects: Application to Facial Analysis

Abstract - *One of the important obstacles in the image-based analysis of the human face is the 3D nature of the problem and the 2D nature of most imaging systems used for biometric applications. Due to this, accuracy is strongly influenced by the viewpoint of the images, being frontal views the most thoroughly studied. However, when fully automatic face analysis systems are designed, capturing frontal-view images cannot be guaranteed. Examples of this situation can be found in surveillance systems, car driver images or whenever there are architectural constraints that prevent from placing a camera frontal to the subject. Taking advantage of the fact that most facial features lie approximately on the same plane, we propose the use of projective geometry across different views. An Active Shape Model constructed with frontal view images can then be directly applied to the segmentation of pictures taken from other viewpoints. The proposed extension demonstrates being significantly more invariant than the standard approach. Validation of the method is presented in 360 images from the AV@CAR database, systematically divided into three different rotations (to both sides), as well as upper and lower views due to nodding. The presented tests are among the largest quantitative results reported to date in face segmentation under varying poses.*

Adapted from F.M. Sukno, J.J. Guerrero and A.F. Frangi. Projective Active Shape Models for Pose-variant Image Analysis of Quasi-Planar Objects: Application to Facial Analysis. *Submitted for publication*

4.1 Introduction

One of the important obstacles in the image-based analysis of the human face is the 3D nature of the problem and the 2D nature of most imaging systems used for biometric applications. Facial images present large changes in shape and appearance when the relative angle between the camera and the face is modified. This three-dimensional nature of the head is further complicated by the non-rigid motion it can involve.

Several works tackle pose variation by learning the relationships between different views. Fan et al. [65] learn a pose change model from example images. A Gaussian skin-color model is used to coarsely detect faces under varying viewpoints and a feature-based strategy provides further refinement and rejection to false alarms. Beymer et al. [14] and Sanderson et al. [166] learn prior information of the face from multiple 2D views of a prototype training set. This allows that, at a later stage, a single view per person would be enough to train a face recognition system. The transformation between different views is handled by optical flow warping in [14], while in [166] the authors used maximum likelihood linear regression (MLLR) and standard multivariate linear regression (LinReg). In the MLLR approach, a generic face model is constructed for each viewpoint (similar to the Universal Background Model used in speech recognition [120]).

Another successful strategy to deal with pose changes has been the use of specific points to construct a graph representation of the face [124, 200]. At each point Gabor features are computed and the recognition is cast as a graph matching problem. These methods account for recognition under varying viewpoint as long as such views are present in the training set. A similar idea was proposed by Maurer et al. [129]. However, in this case the authors propose an interesting transformation between the faces at different poses by assuming that a (small) neighborhood of the nodes is planar. They notice that, under such assumption, the transformation of the jets becomes a purely geometric problem. Their method needs the normals to the graph nodes on the image and an estimation of the rotation angles of the face. They address these problems only partially, by means of a learning strategy from a multiple-pose training set. This constitutes the main drawback of their method.

Another group of approaches can be identified as based on statistical models. As a general rule the models for facial images are bi-dimensional and cannot handle large pose variations. Combining a number of them to extend their viewpoint range has been a popular solution: multiple face templates [13], view-based eigenfaces [137] view-based Active Appearance Models (AAMs) [42,47] and view-based Direct Appearance Models [208] are some examples. This idea is followed also by Li et al. [112] and Xin et al. [203]: the whole range of views from frontal to side views is partitioned to construct separated statistical models. The model to be used for an unknown image is determined with the help of a multi-view face detector [213].

The use of a single statistical model to deal with the whole range of views was

proposed by Romdhani et al [158, 159]. They used KPCA (Kernel Principal Component Analysis) to make the point distribution model non-linear and added the viewing angle as an additional parameter to the landmarks vector. Du and Lin [61] presented a multi-view face synthesis method based on the non-linear generalization of bilinear factorization models, which were trained to stand for two different factors: human identity and head pose. The face description was based on AAMs and, after training with a multi-view database, they were able to generate unseen views from a single 2D image.

Wan et al. [196] propose a different extension of Active Shape Models (ASMs) [44]. They divide the facial shape into two parts: the face contour and the remaining facial features (eyes, brows, mouth and nose), based on the observation that they are not affected in the same way under perspective transformations. The two parts are modeled separately, linking them by a cost function to ensure they still represent a (plausible) human face. In other words, the authors aim at providing more flexibility to the facial shape by softening the link between two parts that behave differently when varying the viewpoint.

The differences pointed out by Wan et al. [196] for behavior of the face contour with respect to the rest of the face can be explained based on projective geometry concepts [80]: the points describing the shape of eyes, brows, nose and mouth are approximately coplanar, while the points of the contour are not [177]. The coplanar approximation was also used by Black et al. [16], combined with affine and curvature models to estimate facial motion based on optical flow.

Projective geometry theory was used by Dias and Buxton [27, 59] to deal with the alignment of shapes under different viewpoints in ASMs. By restricting themselves to affine imaging conditions, the authors propose a method to remove pose variation based on two reference views, appropriately selected from a multi-view dataset. Their Integrated Shape and Pose Model (ISPM) is presented as an extension to the Linear Combination of Views (LCV) [189] under affine conditions. An important point in the work of Dias and Buxton [59] is the selection of a subset of facial landmarks (although manually) for the alignment, based on the observation that the face is not a rigid object and substantial shape differences may be present in the different views to be aligned.

Projective geometry was also used in several works for determining head pose. Wong et al. [197] exploited the vanishing point of the eyes- and mouth-line to derive the 3D pose of the head, assuming a single (but calibrated) view. Ratio and length parameters, however, must be learned from a training set. Gee and Cipolla [70] estimate facial orientation based on knowledge of the individual face geometry. They use the tip of the nose and the four eye-corners, which they assume collinear in 3D. A similar approach can be found in the work of Horprasert et al. [82]. It is interesting how their solution to determine the yaw angle depends only on the relative distances of the four corners of the eyes and the focal length of the camera while for the pitch angle they have to use *typical* anthropometric data.

More sophisticated solutions tackle the problem by dealing with a three-dimensional model of the face. One of the most successful approaches is the Morphable Model [18], in which a large dataset of 3D face scans containing both geometric and textural information are used to construct a 3D face model coined multidimensional morphing function. The main drawback of 3D approaches is related to their computational load and imaging systems considerably more expensive than their 2D counterparts and with some sensitivity issues in the capturing process [153].

There are also some approaches half way between 2D and 3D, which derive 3D shape models from multiple 2D views but perform the image search in 2D. This is the strategy followed by Xiao et al. [202] and Mathews et al. [128], both based on AAMs; and by Li et al. [113] who jointly optimize overall appearance, local appearance (around landmarks), and the difference to the previous frame (when tracking). An interesting point in the combined loss function of Li et al. [113] is the introduction of a *visibility weight* for the appearance of each landmark, which depends on pose (based on the normal to the landmark in the 3D shape). The method was reported to behave reliably in the range of $[-70, 70]$ degrees in yaw. although no quantitative results were provided. Also Tong et al. [186] combine 2D and 3D: the authors assume that a 3D model is available and use it to estimate head pose. Then, the shape model is *corrected* to match the estimated pose using an affine approximation.

A common drawback of all the above techniques is that they need somehow large databases to construct the facial models. Even in [14, 166], where the use of a single training-view per person is investigated, there is the need for a multi-view database from which to learn prior information about pose changes. These databases must be, in general, manually segmented and annotated, which is time consuming, tedious and subjective. On the other hand, frontal-view facial databases are more commonly available and some of them are already annotated [86], [133].

Taking advantage of this, and by means of projective geometry, we replace the 2D similarity transformation relating image and model coordinates in ASMs by a homography-based alignment, which will allow for the segmentation of faces under pose variations based on a model trained only with a database of frontal views. A coplanarity condition of the facial PDM is enforced by excluding landmark points from the contour of the face silhouette and half the nose points. Additionally, as the angles and length rations are not preserved under perspective projection, the sampling locations for the local intensity profiles are computed in model coordinates and then mapped into the image by the homographic transformation. The allowed angles are restricted to views on which both eyes can be seen. As long as this is fulfilled, the method can process poses that include any combination of out-of-plane rotations.

The proposed method can be employed even when the distance from the camera to the face is small and parallelist is lost due to perspective projections. This pose-variant effect can not be captured by affine approximations even if planar objects

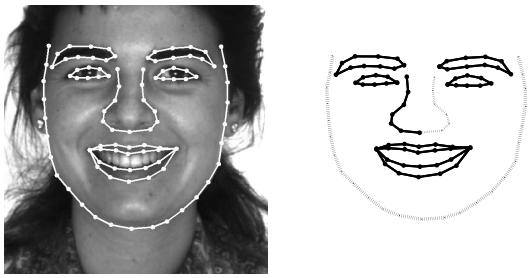

Figure 4.1: Sample image from the AR database [126] with the 98-point annotation template superimposed (left), and the same shape alone (right). The highlighted points are approximately coplanar and can be aligned by homographies through different views [177].

are observed [80]. The projective shape model avoids this difficulty, which makes our approach especially useful for some access control points, images of drivers taken within a car [143], or whenever the face must be captured under constrained and/or variable viewpoints.

We give an overview of the Active Shape Models in Section 4.2 and present our projective extension to it in Section 4.3. Several experiments are presented to demonstrate the proposed technique. Firstly, we investigate the optimal subsets of landmarks for multi-view projective alignment in Section 4.4. In Section 4.5 the projective ASM is used for segmentation and Section 4.6 concludes the paper.

4.2 Active shape models

Active Shape models are based on the combination of a Point Distribution Model (PDM) plus a local image appearance model around each of its points (*landmarks*). The PDM consists in a set of landmarks placed along the edges or contours of the regions to segment. In the case of facial images, these landmarks will be drawn in places like eyes, nose, lips, etc. In our PDM, we use 98 landmarks, distributed over the face as shown in Figure 4.1.

4.2.1 Point distribution model

The PDM is trained by applying Principal Component Analysis (PCA) to the set of landmarked faces. It is generally preceded by a Procrustes alignment [74], in order to make the analysis independent from 2D rotation and scaling (similarity transformations). Trained in this way, the shape variations captured by the PDM correspond to inter-subject shape variations, expression variations and, of course, 3D pose changes, which will not be canceled by the two dimensional alignment.

Let $\mathbf{u}_i$ denote the *i-th* shape (the set of landmarks from the *i-th* face image), and assume there are N_I landmarked shapes of L landmarks each, expressed as x and y coordinates in 2D Euclidean space:

$$\mathbf{u}_i = (x_i^{(1)}, y_i^{(1)}, x_i^{(2)}, y_i^{(2)}, ..., x_i^{(L)}, y_i^{(L)})^T \tag{4.1}$$

$$\overline{\mathbf{u}} = \frac{1}{N_I} \sum_{i=1}^{N_I} \mathbf{u}_i \tag{4.2}$$

$$\mathbf{S} = \frac{1}{N_I - 1} \sum_{i=1}^{N_I} (\mathbf{u}_i - \overline{\mathbf{u}})(\mathbf{u}_i - \overline{\mathbf{u}})^T \tag{4.3}$$

where $\overline{\mathbf{u}}$ is the mean shape and $\mathbf{S}$ the covariance matrix of the set, which is decomposed in its eigenvectors $\mathbf{\Phi}$ and eigenvalues λ_j, for j between 1 and $2L - 1$. Denoting $\mathbf{b}_i$ the PCA-space representation of the shape $\mathbf{u}_i$, they can be related by:

$$\mathbf{b}_i = \mathbf{\Phi}^T(\mathbf{u}_i - \overline{\mathbf{u}}) \tag{4.4}$$

$$\mathbf{u}_i = \overline{\mathbf{u}} + \mathbf{\Phi}\mathbf{b}_i \tag{4.5}$$

It is possible to use only the first M eigenvectors with largest eigenvalues. In that case the above formulas become approximations, with an error depending of the magnitude of the excluded eigenvalues. Furthermore, each component of $\mathbf{b}_i$ is bounded to ensure that only *valid shapes* are represented:

$$\begin{aligned} &|b_i^{(m)}| \leq \beta\sqrt{\lambda_m} \\ &1 \leq i \leq N_I, \quad 1 \leq m \leq M \end{aligned} \tag{4.6}$$

where β is a regularization constant, usually set between 1 and 3, according to the degree of flexibility desired in the shape model.

4.2.2 The intensity model

The intensity model must supply the PDM with candidate landmark points based on the image pixels. It is constructed by computing second order statistics for the normalized image gradients, sampled at each side of the landmarks, perpendicularly to the shape's contour, hereinafter, the profile. In other words the profile is a fixed-size vector of values (in this case pixel intensity values) sampled along the normal to the contour such that the contour passes right through the middle of the normal. Let $\mathbf{g}_j^{(i)}$ be the normalized gradient for landmark *j-th* of the *i-th* training

image. Then:

$$\overline{\mathbf{g}}_j = \frac{1}{N_I}\sum_{i=1}^{N_I}\mathbf{g}_j^{(i)}, \quad 1 \leq j \leq L \tag{4.7}$$

$$\mathbf{\Sigma}_j = \frac{1}{N_I - 1}\sum_{j=1}^{N_I}(\mathbf{g}_j^{(i)} - \overline{\mathbf{g}}_j)(\mathbf{g}_j^{(i)} - \overline{\mathbf{g}}_j)^T \tag{4.8}$$

Note that there is an individual mean vector $\overline{\mathbf{g}}_j$ and covariance matrix $\mathbf{\Sigma}_j$ for each landmark.

During matching, when the model is expected to automatically segment a non-landmarked face image, a number of gradient profiles are sampled over each of the currently estimated landmark positions. The profile with the lowest Mahalanobis distance to the mean profile for that landmark is assumed to be the most-likely landmark location, and serves as updated input for the PDM.

4.2.3 Shape alignment

As stated above, the shapes in the training set must be aligned before applying PCA. Cootes et al. [44] used Procrustes Analysis to this end, thus minimizing the square sum of distances to the mean shape by means of a similarity transformation. At every iteration, each shape is also re-scaled such that $|\mathbf{u}| = 1$.

Therefore, during both the training and the matching processes, the input points expressed in image coordinates (stored in shape $\mathbf{v}_i$) are aligned into a normalized coordinate system (denoted by $\mathbf{u}_i$)

$$\mathbf{u}_i = s\mathbf{R}\mathbf{v}_i + \mathbf{t} \tag{4.9}$$

Here, $\mathbf{R}$ is a 2×2 Euclidean rotation matrix, $\mathbf{t}$ is the translation vector and s is the scale factor, all of them found by aligning $\mathbf{v}_i$ to the mean shape $\overline{\mathbf{u}}$[1].

During the matching process, the image points are projected into this normalized system, the *model coordinate frame*, and the closest valid shape to them ($\widehat{\mathbf{u}}_i$) is obtained from (4.4), (4.5), (4.6). Finally, $\widehat{\mathbf{u}}_i$ is projected back to image coordinates by inverting (4.9).

4.3 Projective ASM

Despite incomplete three dimensional information, given the picture of a person's face, humans have the ability to compose its appearance from other viewpoints. The major problem when trying to mimic the same skill in computer vision is that,

[1]Notice there is some abuse of notation in (4.9) and the transformation must be applied to each (x, y) pair independently.

as the viewpoint departs from the frontal view, some parts of the face become occluded. The other big problem to deal with is the unknown three-dimensional information, which is difficult to estimate from a single image, and will determine the way landmarks will change their relative position across different viewpoints.

The above reasoning can be applied to a wide variety of three dimensional objects and constitutes an important limitation for 2D shape models. However, in some cases the geometry of the studied objects can be assumed approximately planar, thus allowing for some simplifications. It is well known from projective geometry [80] that different views of a plane can be mapped to each other through a 2D homography matrix without the need for three dimensional information of the points. Therefore, if the PDM is constructed with a subset of the landmarks that can be considered coplanar, its points could be mapped to different views by means of a 2D homography. For example, Fig. 4.1 shows the subset of 67 coplanar points from our 98-point template. Notice that the silhouette contour and half of the nose must be excluded (see [177]). With this concepts in mind, a Projective Shape Model is proposed in the remaining of this section.

4.3.1 Projective geometry equations

Given two images from different views of a planar object, they can be related to one another by a projective transformation known as 2D homography. This transformation is performed by means of the 3×3 homogeneous matrix $\mathbf{H}$, which has 8 degrees of freedom (dof). Therefore, at least eight constraints are needed to estimate $\mathbf{H}$, which in 2D-space means four point correspondences between the images. Notice that each landmark originates three equations, but only two of them are linearly independent.

In order to work with projective geometry, homogeneous coordinates must be used to represent the shapes. Then, the $2L$ component vector $\mathbf{u}_i$ becomes a matrix $\mathbf{U}_i$ of size $3\times L$. Since there are no landmarks at infinity, the first two rows of $\mathbf{U}_i$ are set to the 2D x and y components so that the third is the unity:

$$\mathbf{U}_i = \begin{bmatrix} x_i^{(1)} & x_i^{(2)} & \dots & x_i^{(L)} \\ y_i^{(1)} & y_i^{(2)} & \dots & y_i^{(L)} \\ 1 & 1 & \dots & 1 \end{bmatrix} \tag{4.10}$$

The homography estimation can be solved by applying the Direct Linear Transformation (DLT) algorithm [80] to a set of point correspondences between the shapes to align. Similarly to Procrustes alignment, this algorithm minimizes the square sum of distances, but is able to recover any kind of projective transformation. This makes possible the alignment of 3D rotations not contained in the image plane (for coplanar objects).

Let $\mathbf{V_i}$ be the landmarks for an image captured from a generic viewpoint. Its corresponding coordinates $\mathbf{U}_i$ in the model coordinates frame are:

$$\mathbf{U}_i = \mathbf{H}\mathbf{V}_i \tag{4.11}$$

In the DLT algorithm, the rows of $\mathbf{H}$ are split into a nine-component vector $\mathbf{h}$, such that the system resulting from $\mathbf{U}_i \times \mathbf{H}\mathbf{V}_i = 0$ can be written as $\mathbf{A}\mathbf{h} = 0$. The solution is found by solving for the null space of $\mathbf{A}$, which has rank eight [80].

4.3.2 Homography for non-rigid objects

An important aspect when dealing with non-rigid objects is the need for separating shape variations due to intrinsic and extrinsic factors. A clear example is, again, the human face. When aligning two faces captured under different viewpoints, the computation of the homography between them may be affected by their facial expression and identity. This problem motivated firstly Dias and Buxton [59] and recently Tong et al. [186] to align different views from a subset of facial landmarks. Their experiments showed the usefulness of this approach, although they chose the subset manually, mostly justified by the intuitive fact that nose points should not have strong non-rigid motion.

With the help of annotated shapes from facial databases, we propose to determine such a subset from a statistical study. Let $\{\mathbf{H}_i\}_{i=1}^{N_V}$ be a set of homographies representing suitable transformations between different views of the object class under study, which possess non-rigid motion [4, 16]. These objects are represented by a (coplanar) set of landmarks in homogeneous coordinates, as indicated in the previous section.

Additionally, let $\{\mathbf{\Phi_F}, \mathbf{\Lambda_F}\}$ be the eigenvector and eigenvalue matrices obtained by applying PCA to a single-view database (e.g. frontal) containing representative non-rigid deformations of the object class. By randomly sampling this PCA-space, plausible instances of the object class can be generated, say $\{\mathbf{U}_j^{sth}\}_{j=1}^{N_F}$. Hence, $\{\mathbf{V}_{ij}^{sth} = \mathbf{H}_i^{-1}\mathbf{U}_j^{sth}\}$ is a set of $N_V \times N_F$ synthetic instances of the class of objects under study containing shape changes due to non-rigid motion and viewpoint variations (all mixed together).

For each shape $\{\mathbf{V}_{ij}^{sth}\}$ a homography $\hat{\mathbf{H}}_i$ is computed by aligning it to the mean shape of the single-view database, say $\overline{\mathbf{U}}_F$. This homography, however, is computed by using just a subset of the available landmarks. If the chosen subset is not affected by the non-rigid motion, then $\hat{\mathbf{H}}_i$ will be a good estimation of $\mathbf{H}_i$. By repeating this process with different subsets of points, it would be possible to determine the best subset of points to be used for the alignment.

A further refinement can be applied to take into account the localization error intrinsic to each landmark. In fact, the segmentation accuracy of statistical shape

models usually presents strong variations across different landmarks (see [26,178]). Hence, instead of aligning each $\{\mathbf{V}_{ij}^{sth}\}$ to $\overline{\mathbf{U}}_F$, a distorted version $\hat{\mathbf{U}}_F$ should be used:

$$\hat{\mathbf{U}}_F = \overline{\mathbf{U}}_F + \begin{bmatrix} \Delta_x^{(1)} & \Delta_x^{(2)} & \dots & \Delta_x^{(L)} \\ \Delta_y^{(1)} & \Delta_y^{(2)} & \dots & \Delta_y^{(L)} \\ 1 & 1 & \dots & 1 \end{bmatrix} \tag{4.12}$$

where $\Delta_x^{(k)}$ and $\Delta_y^{(k)}$ are the x and y displacements, obtained by randomly sampling the distribution of the localization error estimated for the *k-th* landmark.

4.3.3 Intensity model

The extension of the shape models to work with different views has important effects on the intensity model, since the angles and length ratios between model and image coordinates are not preserved any longer (as they were under similarity alignment). This implies that the normal directions to the contours of the shape will be viewpoint dependent, as well as the spacing among the points sampled for the intensity profiles. This problem can be addressed by generating the sampling locations for the profiles in model coordinates. Subsequently, each point is converted into image coordinates by means of the inverse homographic transformation $\mathbf{H}^{-1}$ and image intensities are then sampled therefrom.

4.4 Experiments on alignment of point subsets

4.4.1 Facial datasets

In this section we address the estimation of the homography between different views due to the non-rigid motion of the face. The problem can be stated as follows: which is the subset of points (among all available facial landmarks) that best estimates the projective transformation between different views?

The method proposed in Section 4.3.2 requires two inputs: a set of homographies $\{\mathbf{H}_i\}_{i=1}^{N_V}$ representing suitable viewpoint transformations, and an eigenspace $\{\mathbf{\Phi_F}, \mathbf{\Lambda_F}\}$ from a single-viewpoint dataset showing representative shape variations from all other factors but viewpoint.

The set of homographies was computed from the AV@CAR database [143]. The variable-viewpoint dataset (VVDS) from this database is composed by 360 manually annotated pictures. For each of 40 users there are nine pictures: one frontal, three left-views, three right-views, one facing up and one facing down.

The 40 *frontal* shapes were used to compute a *frontal PDM*, and all shapes of the VVDS were fitted by this model using homographic alignment (as in [177]). Hence, a set of 360 representative homographies was obtained.

The single viewpoint dataset was constructed by combining 532 shapes from the AR database [126] and 546 shapes from the Equinox database [171], described in [179]. Both databases show frontal shots, with considerable expression variability and only residual viewpoint changes. The resulting dataset, of 1078 shapes, allowed constructing a *frontal* eigenspace with an important amount of variability in facial expression and identity.

4.4.2 Evaluating point subsets

In order to compare the goodness of the different estimations of $\{\mathbf{H}_i\}_{i=1}^{N_V}$ some metric must be adopted. The performance of homographies is generally evaluated with the objective of improving the fitting between (noisy) sets of points [35, 80]. However, our problem is slightly different. The *exact* transformation is known, and its best estimate is not necessarily the one that best fits the image points. An indicative example is showed in Fig. 4.2; the mean shape $\overline{\mathbf{U}}$ was *deformed* by the frontal PDM ($+\mathbf{\Phi}_F\mathbf{b}$) and projected onto a different viewpoint by inverting the synthetic homography $\mathbf{H}$. Then, the estimation $\hat{\mathbf{H}}$ was computed by *projectively aligning* $\mathbf{H}^{-1}(\overline{\mathbf{U}} + \mathbf{\Phi}_F\mathbf{b})$ with $\overline{\mathbf{U}}$. Formally, we applied the DLT algorithm, solving for $\overline{\mathbf{U}} \times \hat{\mathbf{H}}\mathbf{H}^{-1}(\overline{\mathbf{U}} + \mathbf{\Phi}_F\mathbf{b}) = 0$.

The difference between $\mathbf{H}$ and $\hat{\mathbf{H}}$ is illustrated by transforming a *reference rectangle*[2], denoted by $\mathbf{\Pi}$. This rectangle is the bounding box *enclosing* the mean shape (in the frontal-view), and is representative of the *facial plane*. Comparing $\mathbf{\Pi}$ and $\mathbf{H}^{-1}\mathbf{\Pi}$ in Fig. 4.2, it is evident that the transformation $\mathbf{H}$ mainly represents a head rotation to the left. However, since the mouth became opened, the points between $\mathbf{H}^{-1}(\overline{\mathbf{U}} + \mathbf{\Phi}_F\mathbf{b})$ and $\hat{\mathbf{H}}^{-1}\overline{\mathbf{U}}$ fit better by adding a facing-up effect in $\hat{\mathbf{H}}^{-1}$ (so that the mouth is imaged bigger).

The difference between $\mathbf{H}^{-1}\mathbf{\Pi}$ and $\hat{\mathbf{H}}^{-1}\mathbf{\Pi}$ is useful to understand the concept, but it is not appropriate as a quantitative measure, since it suffers the same disadvantages as the *error in one image* [35]. However, this is easily solved by projecting $\hat{\mathbf{H}}^{-1}\mathbf{\Pi}$ back onto the reference view (the one of the mean). The resulting metric, d_{rr}, is then:

$$d_{rr}(\mathbf{H}, \hat{\mathbf{H}}) = d(\mathbf{H}\hat{\mathbf{H}}^{-1}\mathbf{\Pi}, \mathbf{\Pi}) + d(\hat{\mathbf{H}}\mathbf{H}^{-1}\mathbf{\Pi}, \mathbf{\Pi}) \tag{4.13}$$

where $d(\cdot)$ is some point-based distance (e.g. Euclidean), and the second term is added to force $d_{rr}(\mathbf{H}, \hat{\mathbf{H}}) = d_{rr}(\hat{\mathbf{H}}, \mathbf{H})$. Notice that this metric resembles the *symmetric transfer error* [80], but it is adapted to our special needs.

[2] Although the rectangle is used in the explanation for the sake of simplicity, a rectangular grid was actually implemented. This prevents that its corners have a larger influence on the estimation than the central region.

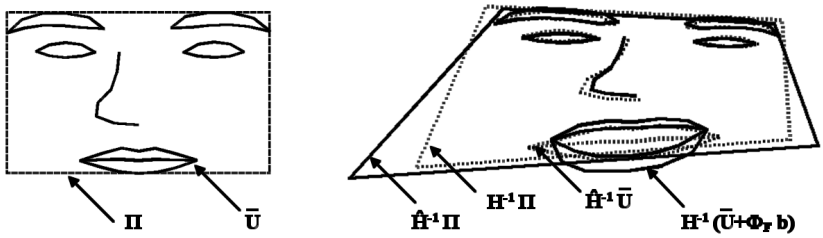

Figure 4.2: On the left, the mean shape $\overline{\mathbf{U}}$ and a *reference rectangle*, $\mathbf{\Pi}$, enclosing it. On the right, a face with the mouth opened under a (synthesized) different viewpoint than the mean. The non-rigid motion makes the estimated homography, $\hat{\mathbf{H}}$, different from the real viewpoint change, $\mathbf{H}$, as it can be seen when transforming the reference rectangle.

4.4.3 Optimal point subset with exact landmark localization

The experiments in this section assume the position of all landmarks are known exactly (noiseless case). That is, $\Delta_x^{(j)} = 0$ and $\Delta_y^{(j)} = 0, \forall j$ in (4.12). The frontal eigenspace described in Section 4.4.1 was randomly sampled (assuming Gaussian distribution according to $\mathbf{\Lambda_F}$) and combined with the homographies estimated from the VVDS. For each homography, a total of 50 random shape instances were generated, resulting in a synthetic dataset of 18,000 shapes. The experiments were split into left-right rotations (14,000 shapes) and up-down nodding (6,000 shapes). Notice that the 40 homographies corresponding to the frontal shot were included in both cases.

As a 2D-homography has 8 dof, the smallest subset of landmarks that can be evaluated contains 4 points. In order to determine the best 4-point subset an exhaustive search was performed among all landmarks in the template ($\simeq$ 766 thousand combinations). Fig. 4.3 and 4.4 illustrate the results of this search. The best four landmarks to estimate left-right head rotation were found to be the two points of the lower lip closest to the corners of the mouth (number 20 and 24 on the template) and two points on the lower contour of the eyebrows. However, there were approximately 30 subsets which achieved very similar performance, without statistically significant difference among them. In fact, the difference of d_{rr} between the first and last rows of Fig. 4.3 was slightly above 3.9%, while the standard error was around $\pm$0.5%.

In spite of the many subsets with similar performance, all of them were composed by two points of the mouth and one of each eye or eyebrow. Regarding the mouth, there was a clear prevalence of points 20 and 24 on the lower lip (see Fig. 4.4), but the two corners of the mouth were also chosen several times within the best 200. As opposed to that, the points on the eyes and eyebrows offered more variability and it was not possible to detect a clear prevalence of any pair of points.

Starting from the best subset of 4 points, the remaining were added one at a time

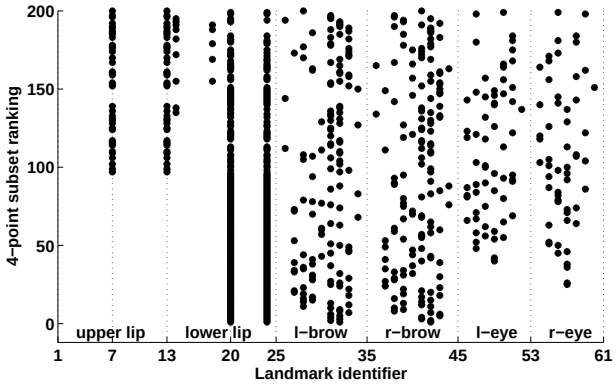

Figure 4.3: The 200 best subsets of 4 landmarks to estimate left-right rotations when the position of each point is assumed to be known exactly (noiseless). Each row shows the 4 selected landmarks, sorted from best to worse according to d_{rr} as defined in (4.13). There is little or no statistical significance for the difference between subsets closer than 30 rows.

by means of a Sequential Forward Selection [100] until all points were included. The obtained average and standard error for d_{rr} are showed in Fig. 4.5. All values were divided by the average d_{rr} obtained when using all landmarks under noiseless conditions for left-right rotation: $\Delta_x^{(j)} = 0$ and $\Delta_y^{(j)} = 0, \forall j$ in (4.12).

It can be seen that using only 7-8 landmarks it was possible to achieve the same performance obtained when using all landmarks. Furthermore, the best estimation of the homographies was achieved for a subset containing between 30 and 40 landmarks (average $d_{rr} \simeq 0.85$ in Fig. 4.5). The *worse* landmarks (the ones not selected within the best subset) belong mainly to the lower lip of the mouth. All other regions contribute to this set in roughly equal proportions among them.

Regarding up-down nodding, the results of selecting the best four landmarks were completely analogous to the experiments for left-right rotations. However, as shown in Fig. 4.5, nodding produced lower errors in the estimation of the homographies than lateral head rotations.

4.4.4 Optimal point subset with uncertain landmark localization

Fig. 4.6 shows the average d_{rr} with and without taking into account the expected localization accuracy for each landmark. This accuracy was modeled from (previous) segmentation experiments: the 1078 images from AR and Equinox datasets described above were segmented using ASM [179]. Comparing the results to manual annotations, the distribution of the localization error was estimated for each

Figure 4.4: Best four points for the estimation of homographies in left-right head rotation (triangles). Other points included within the 200 best subsets of four points are also shown (circles). The different colors indicate the number of times that each point was selected within the best 200 subsets, according to the scale indicated on the right.

landmark. The values for $\Delta_x^{(j)}$ and $\Delta_y^{(j)}$ in (4.12) were obtained by randomly sampling those distributions.

Then, the experiments from the previous section were repeated, but instead of aligning the synthetic shapes to the actual meanshape, $\overline{\mathbf{U}}_F$, a *noisy version* of it was used ($\hat{\mathbf{U}}_F$ in (4.12)). Recall that $\hat{\mathbf{U}}_F$ is not unique, since $\Delta_x^{(j)}$ and $\Delta_y^{(j)}$ were randomly sampled and then each synthetic shape is aligned to the mean with a different set the x and y displacements.

As Fig. 4.6 shows, the inaccurate localization of the points considerably increased the error in the estimation of the homographies. Additionally, the optimal subset included almost all landmarks and there was little benefit in excluding some of them. The best subset of 4 points was the same as the one for the noiseless case. However, there were less combinations of points achieving results close to the best set. Such combinations were again formed by two points of the mouth plus two points of the eyebrows or the eyes.

4.4.5 Conclusions and model initialization

The results presented in the previous sections show that the non-rigid motion of the face can affect the performance of shape alignment. By aligning with just a subset of the landmarks it is possible to improve the results of the alignment, as long as they are precisely located (see Fig. 4.5). In such a case, the optimal alignment was obtained with about 30 landmarks from a template containing 67 points. Furthermore, the errors in the estimation of viewpoint change due to left-right rotations where considerably higher than those due to up-down nodding.

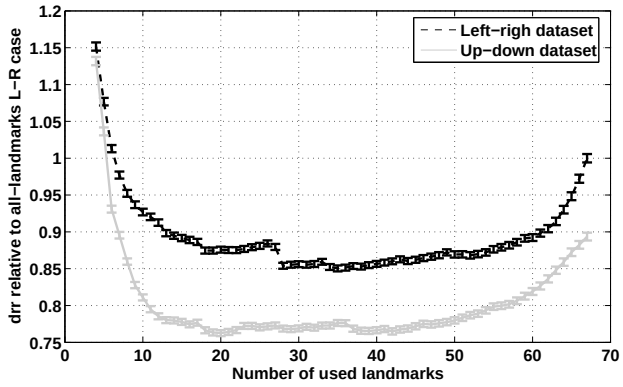

Figure 4.5: Average error in the estimation of homographies for left-right rotations and up-down nodding of the head estimated from different number of landmarks. The average and standard errors of d_{rr} are shown without taking into account the localization accuracy of the landmarks (Section 4.4.3). In both cases, the values are divided by the error achieved when all landmarks are used in the left-right rotation experiment.

If the localization of the landmarks is performed with large uncertainty, then more points are required so that the redundancy compensates for the *noisy* coordinates of the landmarks. From an estimate of the precision expected for the ASM-based segmentation [179], the optimal alignment was obtained by using almost all of the points in the template (see Fig. 4.6). Therefore, the alignment step of the ASM algorithm can be applied to all landmarks without the risk of important drops in performance.

A different problem arises when deciding about initialization. When the model must be applied to a new image, a first guess of the face position is required. When images are captured from a frontal viewpoint, a rectangle roughly enclosing the face is usually enough. But the presence of multiple viewpoints requires more detailed information. One option is the use of feature finders to get the position of certain *key points*. Going back to Fig. 4.5 it can be seen that 7 or 8 points may be enough to achieve the same performance as if all landmarks would be known. But those points are not suitable to be easily localized by feature finders. They are probably well defined in our 67-point template, but do not correspond to anatomical landmarks [60].

To define our initialization set we repeated the experiment of Section 4.4.3 with a reduced set of candidate points, suitable to be located by common feature finders [28, 94, 122, 152, 181, 210]. All the 18,000 shapes were used in a single combined experiment. The 14 candidates are shown in Fig. 4.7, together with the best initialization subset, which consisted in 7 points.

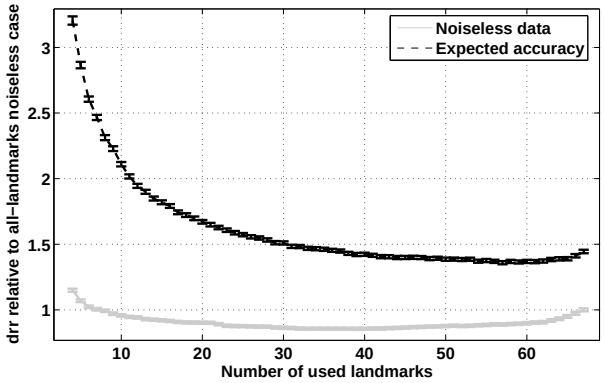

Figure 4.6: Average error in the estimation of homographies for left-right head rotations estimated from different number of landmarks. The average and standard errors of d_{rr} are shown with and without taking into account the localization accuracy of the landmarks (Sections 4.4.3 and 4.4.4, respectively). In both cases, the values are divided by the performance achieved when all landmarks are used without any localization error.

4.5 Experiments on segmentation

In this section we evaluated the influence of the proposed projective alignment in the segmentation accuracy of ASMs. The datasets used to this goal were the same described in Section 4.4.1.

4.5.1 Single-viewpoint model

The main hypothesis of this work is that a homographic alignment allows enhancing the accuracy of ASMs when used to segment different views from those of the training set. To evaluate this, two models were constructed with the 1078 images (and shapes) of the frontal dataset: one using similarity alignment (ASM) and the other using projective alignment (PASM). Then, the models were used to segment the 360 images of the VVDS and results were compared to manual annotations by computing the point-to-curve distance.

Both models were constructed with three resolution levels (to enhance capture range) and by using the same parameters detailed in [179]. During segmentation, all landmarks were used for alignment and both models were initialized with the position of the 7 landmarks selected in Section 4.4.5.

The results of this experiment are shown in Fig. 4.8. For each viewpoint a separate boxplot [130, 193] was computed, including a 95% confidence interval for the median. It can be seen that the PASM had lower median than ASM for all tested

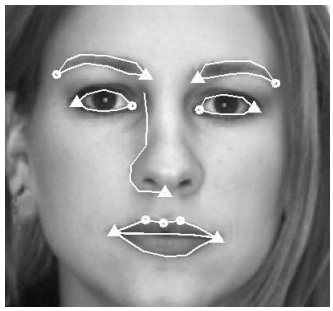

Figure 4.7: Best seven points for the initialization of homographies under head rotations and nodding (triangles), from a reduced set of biological landmarks, suitable to be detected by feature finders. The other candidate landmarks that were not selected are also displayed (circles).

views. This difference tends to grow as the rotation angle of the head increases, becoming statistically significant at 40 degrees and above.

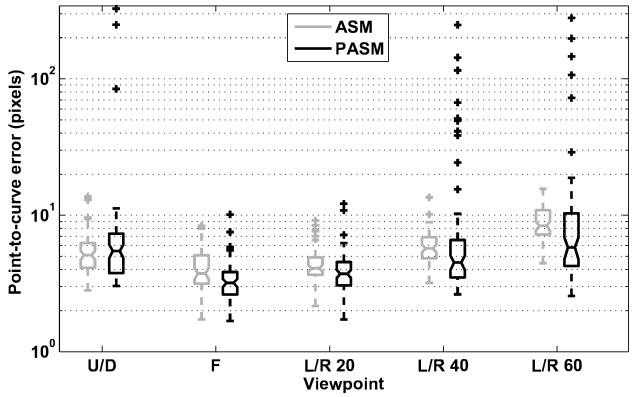

Figure 4.8: Segmentation results on AV@CAR dataset grouped by viewpoint: left-right head rotations (L/R) at approximately 20, 40 and 60 degrees; frontal views (F) and up-down nodding (U/D). Both ASM and PASM models were constructed with a frontal-view dataset from AR and Equinox databases.

On the other hand, PASM produced more outliers, whose segmentation error was considerably higher than the outliers of ASM (note the logarithmic scale of the vertical axis in Fig. 4.8). Indeed, the 8 dof of a homography make PASM more sensitive to mistakes in the image model than ASM, whose similarity transformation has only 4 dof.

An illustrative example is provided in Fig. 4.9. After a few iterations the model

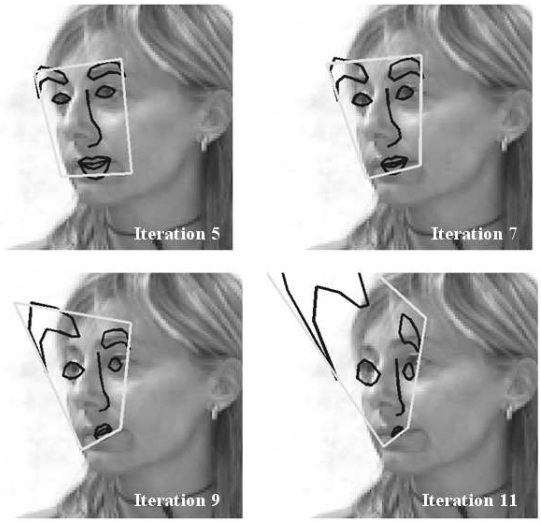

Figure 4.9: Four iterations of the PASM in a diverging example. At each iteration the image is presented with the estimated shape (in black) and reference rectangle (in white) as defined in Section 4.4.2.

has lost track of the right eyebrow and some of its points lie on the white background. Therefore, no suitable candidates can be found (by the image model) for the eyebrow in the next iterations and its position is guided randomly by some background noise or shadows. Under similarity transformations the other face landmarks would prevent the eyebrow to get too far (in the worse case there would be a steady growth of the whole facial size). But with 8 dof the distorted shape of iterations 9 and 11 may be a plausible face due to an extreme camera viewpoint. Hence, that set of points becomes a plausible shape and the model diverges far from the actual face.

4.5.2 Constraining the transformations

A quick analysis of the outliers produced by PASM in Fig. 4.8 shows that many of them could be easily avoided. They are far from the provided initialization and, in many cases, they are even out of the limits of the image. Furthermore, the example of Fig. 4.9 suggests that the clue to the problem is to appropriately estimate the projective transformations.

A simple way to demonstrate this hypothesis is to keep track of previous transformations when estimating the current one. That is, at each iteration the homography relating model coordinates and image coordinates is computed from the current correspondences (4.11) plus all of the previous ones. Hence, at each iteration

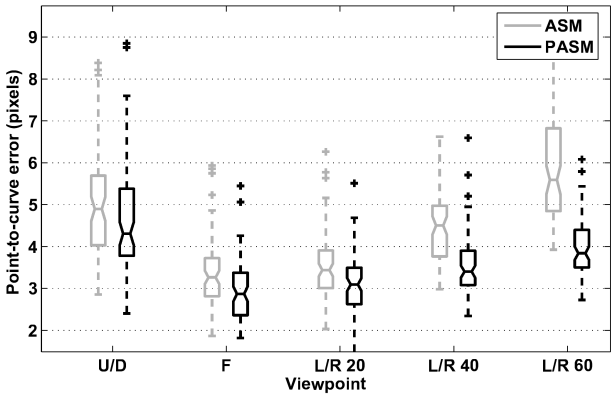

Figure 4.10: Segmentation results with frontal-view models on AV@CAR dataset grouped by viewpoint: left-right head rotations (L/R) at approximately 20, 40 and 60 degrees; frontal views (F) and up-down nodding (U/D). Both ASM and PASM models were aligned by keeping track of all previous correspondences.

the transformation is refined, but not completely re-estimated. This strategy was applied independently at each resolution level, using the result from the previous level as initialization.

The segmentation results by applying this restriction are shown in Fig. 4.10. It is clear that, if initialization can be trusted, this is a suitable approach for PASM to discard most outliers. Additionally, the accuracy of ASM was also improved, but only marginally since the larger limitation in this case resides in the 4 dof of the alignment (i.e. see Fig. 4.11).

Additionally, the results of Fig. 4.10 match the expected behavior for left and right rotations, but are a bit surprising regarding nodding. By analyzing the segmented images it was observed that the curved appearance of the lips under such viewpoints favors an opened mouth shape (while almost no image in the test set showed an opened mouth). Hence, the average segmentation error of the shape gets dominated by the error in segmenting the lips, both in ASM and PASM.

Fig. 4.12 shows the variation of the segmentation error as different number of points are used for initialization. In the case of ASM, two points are enough to initialize the model, while PASM requires four points to estimate the projective transformation. However, it can be seen that the fewer points used, the higher the segmentation error. Moreover, this effect increases with the rotation angle of the head (with respect to the frontal view).

Figure 4.11: Example of a profile view image segmented using models constructed with frontal views only: similarity aligned ASM (Left) and projective ASM (Right).

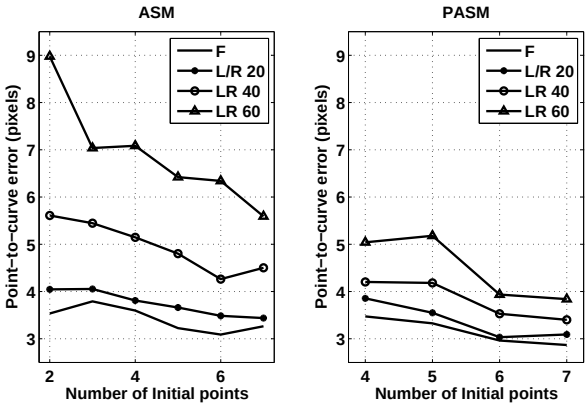

Figure 4.12: Median point-to-curve segmentation error on AV@CAR dataset while varying the number of points used to initialize the models. Separated curves are displayed for the frontal views (F) and different left and right rotation angles of the head (L/R).

4.5.3 Separating the sources of error

The overall segmentation error of ASMs ($\mathbf{e}_{OVL}$) can be separated into two factors: image search error ($\mathbf{e}_{IMS}$) and PDM reconstruction error ($\mathbf{e}_{PDM}$). The first factor is essentially due to limitations of the appearance models of each landmark when locating candidate points on the image. The PDM reconstruction error is due to the constrains of equation (4.6) to the model parameters. It depends on the regularization constant β and the transformation employed to relate image to model coordinates, $T[\cdot]$. We can compute the different errors as follows:

$$\mathbf{e}_{OVL} = \hat{\mathbf{v}}_i - \mathbf{v}_i \tag{4.14}$$

$$\mathbf{e}_{PDM} = \mathbf{v}_i - T^{-1}[\overline{\mathbf{u}} + \Phi \mathbf{b}_i] \tag{4.15}$$

$$\mathbf{e}_{IMS} = \hat{\mathbf{v}}_i - T^{-1}[\overline{\mathbf{u}} + \Phi \mathbf{b}_i] \tag{4.16}$$

where $\mathbf{v}_i$ is the ground-truth shape for the *i-th* image, $\hat{\mathbf{v}}_i$ is the shape estimated by the model for the *i-th* image (both in image coordinates), and $\overline{\mathbf{u}} + \Phi\mathbf{b}_i$ is the best PDM reconstruction for $T[\mathbf{v}_i]$ (under the constraints of (4.6)), Notice that $\mathbf{e}_{IMS}$ may also be written as:

$$\mathbf{e}_{IMS} = T^{-1}[\overline{\mathbf{u}} + \Phi\hat{\mathbf{b}}_i] - T^{-1}[\overline{\mathbf{u}} + \Phi\mathbf{b}_i] \tag{4.17}$$

being $\overline{\mathbf{u}} + \Phi\hat{\mathbf{b}}_i$ the PDM reconstruction for $T[\hat{\mathbf{v}}_i]$, which is an exact representation since $\hat{\mathbf{v}}_i$ has been generated by the model. Hence, $\mathbf{e}_{IMS}$ is the *deviation* of the model parameters due to imprecise image search but it is not independent of $\mathbf{e}_{PDM}$. The overall error is the sum of vectors $\mathbf{e}_{IMS}$ and $\mathbf{e}_{PDM}$, whose relative orientation is not evident.

Fig. 4.13 shows the three errors for ASM and PASM for the different viewpoints of the VVDS. It is evident that, as the head rotates to the left or to the right, the magnitude of $\mathbf{e}_{PDM}$ considerably increases, while $\mathbf{e}_{IMS}$ remains roughly the same (especially for PASM). The use of a projective transformation does not avoid the increase of $\mathbf{e}_{PDM}$ but considerably limits it: for frontal views, the magnitudes of $\mathbf{e}_{PDM}$ for ASM and PASM differ about 10% while for 60 degrees of left-right rotation (L/R 60) the difference becomes greater than 50%.

Fig. 4.13 also allows to get some insight on the high segmentation errors for up/down head nodding. Although the PDM reconstruction errors are not as high as those for L/R 60, the overall error for the nodding views is dominated by the image search error, which is the highest of all views.

It is important to emphasize that all errors showed in Fig. 4.13 depend on the regularization constant β. If β is too small the PDM has not enough flexibility to represent plausible shape variations, increasing $|\mathbf{e}_{PDM}|$. For β too high the PDM can also represent non-plausible shapes, which results in less constraints for the image search and increases $|\mathbf{e}_{IMS}|$ (see also [177]). For example, Fig. 4.9 can be seen as an extreme example of such situation. The lowest segmentation error is achieved for some *compromise* value of β, usually between 1 and 3, that we have set to 1.5 based on [179].

4.5.4 Multi-viewpoint model

We evaluate the case when a multi-view training set is available. To this goal the VVDS was divided into halves (each composed of all available views for 20 people) and 2-fold cross validation was performed. Therefore, the training and test sets were disjoint regarding identities, but shared the spanned viewpoints. Notice that, as symmetry was exploited to use the same template for left and right rotations (by mirroring it), the total rotation range covered by the models was about 60 degrees.

The results are displayed in Fig. 4.14. Both ASM and PASM exhibit a small performance variation between frontal view and rotations up to 40 degrees. However,

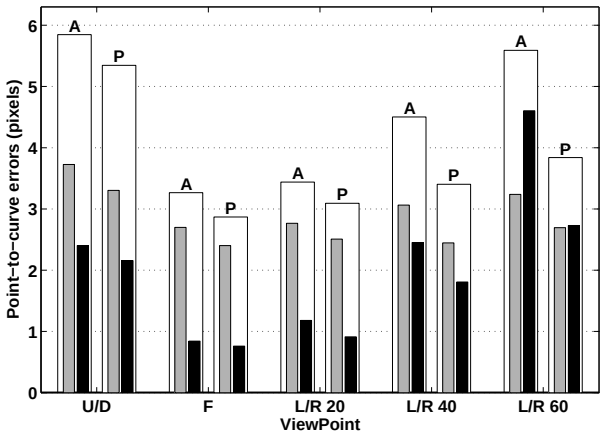

Figure 4.13: Median point-to-curve errors on AV@CAR dataset using frontal-view models for ASM (A) and PASM (P) and grouped by viewpoint: left-right head rotations (L/R) at approximately 20, 40 and 60 degrees; frontal views (F) and up-down nodding (U/D). The bars indicate the magnitude of three different errors: the overall segmentation error (white), the image search error (gray) and the PDM reconstruction error (black).

the segmentation error for ASM grows considerably for views rotated 60 degrees, and both models get strongly affected by up and down nodding.

In order to interpret these results, it is useful to look at the main modes of variation of both models, which are shown in Fig. 4.15 and 4.16. The main variation of the training set is left-right rotation, which is captured by the first mode of variation in ASM. The projective alignment effectively deals with the left-right rotations, hence such a variation is not present in the modes of variation of PASM.

On the other hand, the second mode of ASM is very similar to the first mode of PASM, indicating that important nodding effects have not been successfully removed by the projective model. Again, the curvature effects observed on the lips and eyebrows may be responsible of this behavior.

4.5.5 Discussion

It is interesting to compare the results of Sections 4.5.2 and 4.5.4, mainly shown in Fig. 4.10 and 4.14, respectively. Both plots indicate the segmentation errors of ASM and PASM on images with different viewpoint. However, in Section 4.5.2 the models were constructed exclusively with frontal-view images, while in Section 4.5.4 all test views were also available for training. Fig. 4.17 shows some examples from the AV@CAR database matched with ASM and PASM using frontal-only models and

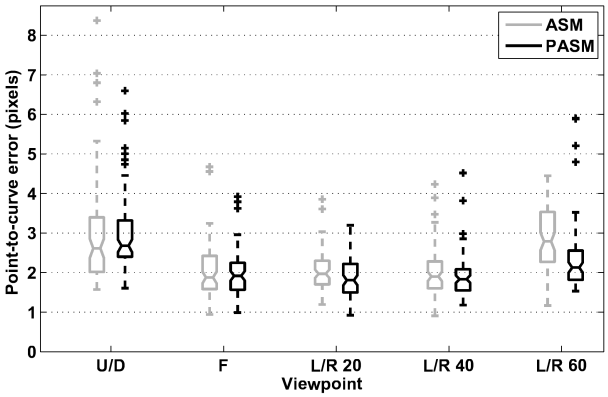

Figure 4.14: Segmentation results from 2-fold cross validation on AV@CAR dataset grouped by viewpoint: left-right head rotations (L/R) at approximately 20, 40 and 60 degrees; frontal views (F) and up-down nodding (U/D).

multi-view models.

Both ASM and PASM are considerably more robust to left and right head rotations when a multi-view dataset is available. In the case of ASM this matches the expected behavior: the similarity alignment cannot cope with out-of-plane rotations; hence if the PDM was constructed only with frontal images, the other views will have a higher reconstruction error, as clearly shown in Fig. 4.13.

In the case of PASM, where a projective transformation is employed, one may expect to get the same results with frontal models and multi-view models. However, this is not the case, and the reason is the coplanarity assumption made about our face model, which is only an approximation. When the model is trained only with frontal views, the PASM finds the best *coplanar* representation for the given viewpoint but the surface of the face, however, is slightly curved. Hence, not all points lie exactly in the same plane, which can be observed especially in the eyes, the mouth and the eyebrows, whose outer corners usually lie *slightly behind* the rest of the points in our PDM.

When the model is trained with multiple views, those curvature effects are learnt by the PDM and complement the projective transformation. This effect is especially clear in rows 2, 3 and 5 of Fig. 4.17 (when comparing columns 2 and 4). Nonetheless, in practice all frontal-view databases do have some (slight) pose changes [73] which allow the PDM to capture some of the curvature effects not covered by the projectivity.

Finally, note that the difference between ASM and PASM is considerably larger for frontal models than it is for multi-view models. And in the latter case this difference increases as the viewpoint departs from frontal, consistently with results

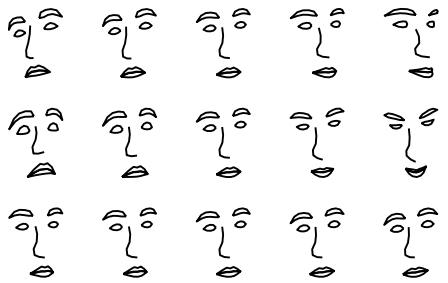

Figure 4.15: From top to bottom, the first three modes of variation of ASM trained with multiple views. From left to right, the variation due to -2, -1, 0, 1 and 2 times the standard deviation of the corresponding mode.

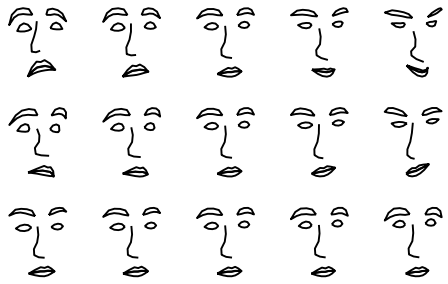

Figure 4.16: From top to bottom, the first three modes of variation of PASM trained with multiple views. From left to right, the variation due to -2, -1, 0, 1 and 2 times the standard deviation of the corresponding mode.

previously reported for ASM [112,158].

4.5.6 Comparison to related work

Comparison of these results to those of related work in segmentation of multiple view faces is a complicated task. There are mainly two problems that hamper such comparison: the lack of annotated data and the absence of a standard evaluation strategy.

Regarding databases, the only annotated and publicly available database to date is the IMM database [139]. It contains 6 images for each of 40 subjects, landmarked with a 58-point template. Two of the images show systematic viewpoint variation (left and right partial profiles, mostly within ± 30 degrees), but the angles of rotation

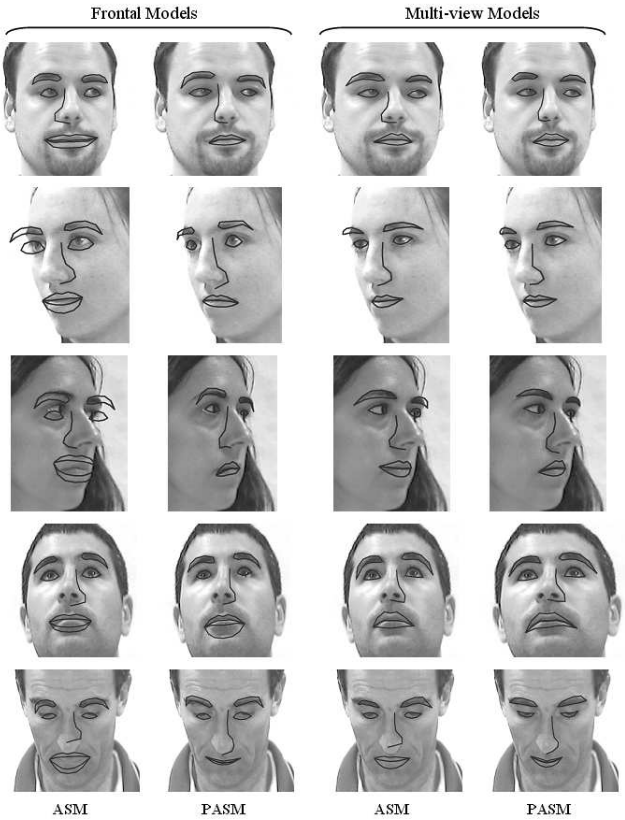

Figure 4.17: Examples of segmentation results. Each row shows a different image segmented with four different strategies: ASM and PASM constructed only with frontal view images (first two columns); ASM and PASM constructed with multiple view images (3rd and 4th columns).

were not controlled and differ significantly among the different individuals. This aspect is linked to the second problem (evaluation) and makes the IMM database not appropriate for our needs. As it was stated by Black et al. [15] testing algorithms with image sets that include only a few qualitative labels for pose angles (such as frontal or 3/4 profile view) is of limited value in assessing how many degrees of viewpoint variation the algorithms can tolerate.

Based on the above statement, published results from which the viewpoint effect on segmentation can be quantified has been gathered in Table 4.1. For each algorithm, the table shows the ratio between the performance for left and right rotated views with respect to the one for frontal views (e.g. within ± 10 degrees). This automatically discards works on which segmentation was demonstrated only visu-

Table 4.1: Comparison of segmentation performance for left and right head rotations with respect to frontal views

Method	Database (# images)	Metric	Rotation range (degrees)			
			11-30	31-50	51-70	> 70
ASM	AV@CAR (280)	P2CSE (average)	1.01	1.01	1.41	n/a
PASM	AV@CAR (280)	P2CSE (average)	0.95	0.96	1.18	n/a
Romdhani et al. [158]	own data (114)	P2PSE (typical)	1.2	1.0	1.2	2.5
Wan et al. [196]	ORL (270)	P2PSE (average)	1.17	n/a	n/a	n/a
Buxton et al. [27]	own data (65)	P2PRE (average)	1.1	n/a	n/a	n/a

ally [47,113,202], those reporting performance on just a few selected examples [159] or the ones not reporting independent results for frontal and profile views [65,216]. Nodding results were not compared since, to the best of our knowledge, no results are available but the ones on this paper.

The results for the first two algorithms of the table are those of this paper. They correspond to the mean point-to-curve segmentation error (P2CSE) from the plots in Fig. 4.14. It can be seen that the error for PASM increases at most 20% with respect to frontal views, while the error for ASM grows up to 40%. On the other hand, both methods demonstrate small performance degradation for rotations under 50 degrees.

Next one listed is the non-linear ASM of Romdhani et al. The values on the table were inferred from the plots of [158], in which a 19-viewpoint dataset was used, ranging from -90 to 90 degrees (left-right rotations). The results were provided as *typical* point-to-point segmentation errors (P2PSE) on a quite small database (only 6 people). This approach exhibits higher variation than ASM and PASM for small rotations, while for rotations between 30 and 70 degrees its performance seems comparable to PASM. It is the only one on the table that can handle full profile views (although its error gets considerable higher).

Wan et al. [196] reported segmentation errors on images from the ORL database [165]. A genetic-based ASM was shown to largely outperform the standard ASM. However, under pose variations, the error of the new method increased (in percentage) as much as the one of standard ASM. Additionally, the test data exhibits small rotations, mostly within ± 20 degrees [185] and the angles are not uniform among individuals.

Table 4.2: Comparison of segmentation performance for left and right head rotations training the models with frontal views

Method	Database Dataset	Metric	Rotation range (degrees)		
			11-30	31-50	51-70
PASM	280 img. (AV@CAR)	P2CSE (average)	1.07	1.20	1.37
ASM	280 img. (AV@CAR)	P2CSE (average)	1.05	1.32	1.74

Multiple models accounting for different viewpoint intervals were tested by Xin et al. [203]. They provided plots of the P2PSE distribution from which the median could be estimated. However, they used different normalization strategies for the errors of frontal- and profile-view images, and the datasets were too poorly described to enable for a correction factor. Furthermore, Xin et al. used separated models for each view, hence the tolerance of the model to pose changes is not really addressed. The same happens in the work by Cootes et al. [46], where they report point-to-point errors for frontal- and profile-view AAMs. The models do not actually handle different viewpoints: instead, there are two single-view models and the goal is to show the advantage of coupling them for the simultaneous segmentation in multiple views.

The results in the last row of the table were obtained from the plots of Buxton et al. [27]. They tested their affine-based ASM on a dataset containing 5 facial expressions and 13 viewpoints, though from just one individual. The metric reported was the average point-to-point *reconstruction* error (P2PRE), meaning that only geometry reconstruction (but not image search) was addressed.

From the collected data, PASM is among the least dependant algorithms with respect to pose. Nonetheless, all results are from different datasets, with different initialization strategies and annotation templates, rendering difficult a conclusive comparison.

It can be seen that the comparative table is small and quite sparse. Actually, the comparison reported in this work (on 280-360 images) is among the largest quantitative evaluations performed on segmentation under systematic pose changes. We believe that one of the major reasons for this shortage is the considerable effort required on annotating multi-view databases.

Some researchers have tried to circumvent this issue by using semi-automatic methods to annotate the data. In spite of some concerns on how accurate are such evaluations, they allow for experiments on larger datasets. The most relevant examples are probably from Mathews et al. [128,202] who reported results on 900 images, but their data consisted on video sequences from only 6 people and pose changes

were not acquired nor organized in a systematic way, thus making it impossible to incorporate their results in Table 4.1.

Differently from all other approaches, PASM can be employed without a multi-view training set. Table 4.2 gathers our results from ASM and PASM when trained with only frontal-view data (Section 4.5.1). The influence of pose variations in the performance of PASM (in terms of P2CSE) is approximately half of that exhibited by ASM.

4.6 Summary and conclusions

In this paper we have presented a projective extension of ASMs, dealing with facial images taken under wide viewpoint variations. The method assumes a coplanar point distribution model and, as far as this requirement is fulfilled, it can be applied to different types of objects.

As pose variation in faces can contain non-rigid motion, a study was performed to determine which points should be used for pose alignment. It was shown that using all the landmarks available in the template is not always the best choice: the same performance can be obtained with less than 10 landmarks, and a 15% improvement is possible by using a 30-point subset. When the error in the localization of landmarks was considered, the advantage of excluding some points vanished out. However, a more accurate localization method may benefit from using these subsets.

The segmentation performance was tested on 360 images taken systematically under different viewpoints. When a multi-view training set was considered, the proposed algorithm demonstrated being almost invariant to head rotations to both sides up to 60 degrees, and comparable to the few published results available in the literature. This paper is the first in quantitatively analyzing nodding variations, which unfortunately lead to poorer performance when compared to left-right rotations.

When only frontal images were considered for training, the proposed method increased its segmentation error up to 30% under head rotations, approximately half of what ASM increased under the same circumstances. In this way, the projective alignment was shown to significantly reduce the variance in performance produced by head pose.

CHAPTER 5

Reliability Estimation for Statistical Shape Models

Abstract - *One of the drawbacks of statistical shape models is their occasional failure to converge. Although visually this fact is usually easy to recognize there is no an automatic way to detect it. In this work we introduce a generic reliability measure for statistical shape models. It is based on a probabilistic framework and uses information extracted by the model itself during the matching process. The proposed method was validated with two variants of Active Shape Models in the context facial image analysis. Experimental results on more than 3700 facial images showed a high degree of correlation between the segmentation accuracy and the estimated reliability metric.*

Adapted from F.M. Sukno and A.F. Frangi. Reliability Estimation for Statistical Shape Models. *Conditionally accepted for publication in IEEE Transactions on Image Processing, pending minor revision*

5.1 Introduction

Since their publication more than a decade ago, Active Shape Models (ASMs) [44] have attracted considerable attention of the research community. ASMs are based on landmark points, whose shape and local intensity distributions are learnt from a *training set*, that is, an already segmented set of images of the class of objects to be segmented. This way, ASMs are able to deal with very different types of objects, just by choosing an appropriate set to train them with. Some examples can be found in [97], [44] or [190], just to mention a few, which analyze people's silhouettes, hands and human hearts, respectively.

Several extensions have been proposed to the original method. Improvements to the image model [191, 206], handling of out-of-plane rotations [59, 158], three-dimensional modeling [53, 142], and spatio-temporal extensions [78] among many others. Active Appearance Models (AAMs) [37] can also be seen as an extension of ASMs, with enough success to have gained independent entity, and their own further extensions.

One of the drawbacks of ASMs, however, is their occasional failure to converge. This situation is illustrated in Fig. 5.1: the shape model was supposed to locate the contours of the face, but it has not succeeded. Although visually this fact is very clear, the method itself does not provide an automatic way to detect such failure. In other words, it does not provide any *reliability measure* of the result. In fact, despite of some attempts to define smarter strategies [45,109,204], the segmentation process is usually stopped by imposing a maximum number of iterations rather than by a convergence criterion [192]. The situation changes in AAMs, since they seek for the convergence of the represented texture, but they can still converge to a wrong result [42].

This is especially important when ASMs are intended to be used into fully automatic systems. For example, face recognition applications such as [91, 106] rely on the segmentation of facial features to determine identity. Results like the ones showed in Fig. 5.1 hamper any recognition attempt, and avoiding (or at least automatically detecting) this kind of situations may be of great advantage.

Some (partial) solutions have been proposed in the literature, mainly focused on specific applications. Wan et al. [196] replace the traditional matching process of ASM by a genetic selection based on an *ad-hoc* fitness function. This function is a weighted sum based on edge intensities and gray-level appearance on the target image, used to validate each of the possible shapes generated by the genetic algorithm. Li et al. [109] use a number of thresholds to discard incorrectly matched points from the shape optimization by defining a measure of shape change as the average change of the landmark positions. They declare convergence when the shape change is below some threshold for at least five iterations in order to handle the oscillatory behavior exhibited by this measure. The proposed method appears to improve the stability of the matching process, although for some cases the num-

Figure 5.1: Three examples of ASM matches to a face image. Two of them (in dark color) are clearly wrong (e.g. due to a bad initialization).

ber of iterations is far over 100, indicating the difficulty to converge. And, most importantly, convergence does not guarantee an accurate result.

On the other hand, unreliable outputs are not a problem specific to ASMs nor to facial biometrics. Automatic measures of reliability have been investigated since a few years for classifier fusion in multi-modal biometrics [160]. In that context, it has been related to the degree of trust offered by a unimodal classifier decision [32, 99]. Thus, reliability measures are used to dynamically weight the outputs from the different modalities and enhance the performance of the multi-modal system. Several metrics have been proposed to measure reliability [64, 156], but mostly focused on classification tasks, and none focused on statistical shape models.

In this work we introduce a reliability measure based on a probabilistic framework [176]. It uses information extracted by the model itself during the matching process (as opposed to [196] in which additional image processing is required) and it can be applied to statistical shape models in general. Its goal is that, for each segmented image, the matching process provides two different outputs: the segmentation result (the estimated landmarks' position) and a confidence measure for the segmentation result. We demonstrate its usefulness with ASMs [44] and IOF-ASMs [179] using four different face databases: XM2VTS [133], AR [126], Equinox [171] and AV@CAR [143].

5.2 Active shape models

This section presents an overview of the Active Shape Model (ASM) of Cootes et al. [44] and the Invariant Optimal Features extension (IOF-ASM) [179]. The explanation is kept parallel for both methods emphasizing the differences when appropriate.

5.2.1 Training process

During the training process a set of annotated images (the *training set*) is analyzed. The annotations consist on a set of points (*landmarks*) defining the contours of interest in the image. The statistical model must learn how to automatically locate this group of landmarks (hereinafter *the shape*). Shape statistics are learnt by means of Principal Component Analysis (PCA) of the landmarks in the training set. On the other hand, the statistics for the appearance are learnt locally for each landmark. Therefore, both ASM and IOF-ASM have one shape model or PDM (Point Distribution Model) and as many appearance models as the number of landmarks composing the shape.

Each appearance model of ASM is constructed by computing second order statistics for the normalized image gradient, usually known as *the profile*: a fixed-size vector of values sampled along the perpendicular to the contour such that the contour passes right through the middle of the perpendicular.

In IOF-ASM each appearance model can be thought as a texture classifier plus a robust decision block. The inputs to the classifier are features based on image derivatives computed in the neighborhood of the landmark. Those image derivatives are appropriately combined to generate differential invariants to rigid transformations, giving the name to the method (for differential invariants please refer to [67,167]). Once trained, the texture classifier receives image features and returns the distance to the most likely position for the landmark. The robust decision block does not require any training and it is explained below.

5.2.2 Model-to-image adaptation process

When the shape models are used for segmentation, only two inputs are required: an image containing a face and a starting guess of the face position (e.g. provided by a face detector). The process begins by placing an average shape at the initial position on the image. Subsequently, at each iteration and for each landmark, the corresponding appearance model is asked about the best position to place the landmark (there are only a few allowed positions to move the landmark, determined by its previous position and the search range the image model is assigned). Then, the landmarks are constrained by the PDM so that they generate a plausible shape. A predefined number of iterations are executed after which the model is assumed to be fitted.

The criterion used by the appearance models of ASM is the minimization of the Mahalanobis distance based to the Gaussian model learnt during training.

In IOF-ASM, instead, the texture classifiers were trained to estimate the distance to the correct landmark position. Hence, when the values analyzed over the candidate positions are plotted we should obtain a “V"-shaped profile. Fig. 5.2 schematically illustrates this: there are seven candidate positions (including the current one)

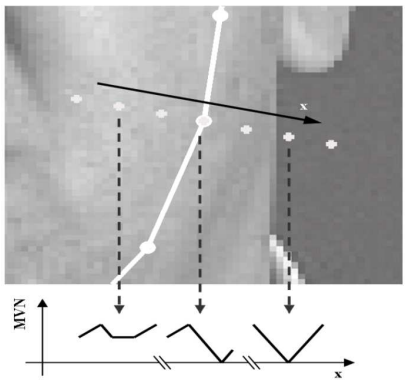

Figure 5.2: Snapshot of the iterative segmentation process for IOF-ASM. The (upper) image shows part of a face with the current fitting of the model (solid line) and the candidate positions to displace one of the landmarks. The outputs from the texture classifier are displayed below, for three of those positions.

and, as the evaluated position departs from the face boundary, the profile shape deforms more and more with respect to the ideal "V".

After the texture classifier evaluated each position, the *robust decision block* determines which one best matches the expected "V", and whether the coincidence is good enough to trust the choice. Specifically, at least two thirds of the profile points supporting one candidate position are required in order to validate the choice. In the cases when all candidates are too far from the V profile, instead of moving the landmark to the *least worse* position, it is kept unchanged (see [179] for details).

5.3 Estimating the reliability of the segmentation

Either due to a wrong initialization or simply because of a segmentation failure, there are cases in which the statistical shape models produce quite incorrect segmentations. For example, in [42], the segmentation of 400 facial images failed to converge for 1% and 1.6% of the test data, using ASM and AAM respectively[1]. The problem is that, if no manual annotations are available, the mode-to-image adaptation algorithms for statistical shape models cannot, in general, tell whether the segmentation results are reliable or not.

The IOF-ASM provides a straightforward way of estimating the reliability of the obtained segmentation: the *robust decision block* of the landmarks' appearance models. As it was explained in Section 5.2.2, these blocks determine which is the

[1]The failure was declared when the average distance from the automatic segmentation to the hand-annotated landmarks was greater than 10 pixels.

best position to displace each landmark, and if that position is good enough (*reliable*) or not. In other words, they tell if the appearance of the image matches the expected appearance learnt for each landmark from the training set. In the next section we show that this idea can be extended to shape models in general.

If the obtained segmentation is accurate enough, most of the landmarks are close to their correct positions on the image. Therefore, in this case, the appearance-based search should be reliable for most landmarks.

5.3.1 Defining reliability

Let I_i be the *i-th* image in the training set, $\mathbf{u}_i$ its associated shape and $\hat{\mathbf{u}}_i$ the (automatic) fitting obtained for image I_i. Each shape is composed by the concatenation of the x and y coordinates of L landmarks:

$$\mathbf{u}_i = (x_i^{(1)}, y_i^{(1)}, x_i^{(2)}, y_i^{(2)}, \quad ... \quad , x_i^{(L)}, y_i^{(L)}) \tag{5.1}$$

$$\hat{\mathbf{u}}_i = (\hat{x}_i^{(1)}, \hat{y}_i^{(1)}, \hat{x}_i^{(2)}, \hat{y}_i^{(2)}, \quad ... \quad , \hat{x}_i^{(L)}, \hat{y}_i^{(L)}) \tag{5.2}$$

Since I_i is in the training set, there is a set of landmarks manually annotated for it that provides the *ground-truth*. The segmentation error for the *j-th* landmark of $\hat{\mathbf{u}}_i$ is denoted as $\epsilon_i^{(j)}$, and the averaged segmentation error of the whole shape as $\mathcal{E}(\hat{\mathbf{u}}_i)$:

$$\epsilon_i^{(j)} = d\left(\{x_i^{(j)}, y_i^{(j)}\}, \{\hat{x}_i^{(j)}, \hat{y}_i^{(j)}\} \right) \tag{5.3}$$

$$\mathcal{E}(\hat{\mathbf{u}}_i) = \frac{1}{L} \sum_{j=1}^{L} \epsilon_i^{(j)} \tag{5.4}$$

where $d(\mathbf{a}_1, \mathbf{a}_2)$ is a distance measure between $\mathbf{a}_1$ and $\mathbf{a}_2$.

The appearance model for each landmark is assumed to provide a binary variable $\hat{r}_i^{(j)}$ indicating whether its fitting was estimated as reliable ($\hat{r} = 1$) or not ($\hat{r} = 0$). The reliability $\hat{r}_i^{(j)}$ can be thought of as an estimation of the random variable $r_i^{(j)}$, which indicates whether landmark j of shape i is correctly placed (i.e. its error is below a certain threshold):

$$r_i^{(j)} = 1(\epsilon_i^{(j)} < \epsilon_{th}^{(j)}) \tag{5.5}$$

where $1(b)$ takes one if b is true and zero otherwise, In other words, we will say that a landmark position is correct or *accurate*, when its distance with respect to the ground-truth position is smaller than a threshold. The term *reliable* will be employed to indicate the estimation performed by the image model.

An important component of the above equations are the values for $\epsilon_{th}^{(j)}$. These thresholds determine whether the landmark positions are considered correct or not.

An interesting criteria to define this is proposed in [125], based on the precision of the method: by computing the segmentation error in the training set, the author estimates the expected value of the method's error. This value is the *intrinsic precision* of the algorithm, so it can be used as a threshold to determine its success or failure in individual cases.

As opposed to [125], in which a global threshold was used, we compute a separate threshold for each landmark, as statistical shape models usually provide different localization accuracy for each of the landmarks [26]. Formally:

$$\epsilon_{th}^{(j)} = \frac{1}{N_I} \sum_{\forall i} \epsilon_i^{(j)} \tag{5.6}$$

being N_I the number of images in the training set.

Notice that, as ASM-based segmentation is an iterative process, the above equations have an implicit dependency on the iteration number, t. Hence, we should write: $\hat{\mathbf{u}}_i(t)$, $\epsilon_i^{(j)}(t)$, $\mathcal{E}(\hat{\mathbf{u}}_i; t)$ and $r_i^{(j)}(t)$, $\hat{r}_i^{(j)}(t)$. For the sake of clarity, this dependency will be omitted until Section 5.3.6, with the following exception:

$$\epsilon_{th}^{(j)} = \frac{1}{N_I} \sum_{\forall i} \epsilon_i^{(j)}(N_T) \tag{5.7}$$

as the precision of the algorithm should be computed at the final iteration, $t = N_T$.

5.3.2 Estimating reliability from the appearance model

In the above formulation the appearance model of each landmark is asked to provide an estimation of the reliability of the current placement (for that landmark).

In general, the appearance of statistical shape models is evaluated according to a certain distance, or *dissimilarity* metric. For example, in the seminal ASM of Cootes et al. [44] the Mahalanobis distance of the intensity gradient is employed. Let $\zeta_i^{(j)}$ be the dissimilarity calculated by the appearance model for landmark j of shape i. We define

$$\hat{r}_i^{(j)} = 1(\zeta_i^{(j)} < \zeta_{th}^{(j)}) \tag{5.8}$$

where the threshold ζ_{th} is chosen to maximize the mutual information between $\hat{r}_i^{(j)}$ and $r_i^{(j)}$ across the training set:

$$\zeta_{th}^{(j)} = \underset{x}{\operatorname{argmax}} \; MI\Big(1(\epsilon_i^{(j)} < \epsilon_{th}^{(j)}); 1(\zeta_i^{(j)} < x)\Big) \tag{5.9}$$

where $MI(x, y)$ is the mutual information between x and y. This is a natural choice, since we seek for a variable $\hat{r}_i^{(j)}$ that provides information about $r_i^{(j)}$ based on appearance data [88].

5.3.3 Reliability of a whole shape

Analogously to the reliability for each landmark, we can define the reliability of shape $\hat{\mathbf{u}}_i$ as a whole. We denote it as $R(\hat{\mathbf{u}}_i)$, or simply R when it refers to a generic shape and clarity is not compromised. To decide whether a shape is reliable or not, we compute the segmentation error averaged for all its landmarks, and compare it with the accuracy expected from the segmentation algorithm, so that each shape can be classified as:

* `reliable` ($R = 1$): when it has reached the accuracy expected from the algorithm (below the threshold $\mathcal{E}_{th}$).
* `unreliable` ($R = -1$): when its error is unacceptably high compared to the accuracy expected from the algorithm (above a second threshold, $\mathcal{E}_M$).
* `undefined` ($R = 0$): when it does not meet either of previous criteria. The segmentation error falls within $(\mathcal{E}_{th}, \mathcal{E}_M)$.

Following [125], the definition of the threshold $\mathcal{E}_{th}$ is evidently the average of all $\epsilon_{th}^{(j)}$. On the other hand, $\mathcal{E}_M$ stands for cases when the error is far above $\mathcal{E}_{th}$ and the algorithm has therefore failed in performing an acceptable segmentation. To set this threshold we adopt the outlier criterion of [68]:

$$\mathcal{E}_M = q_3 + 2(q_3 - q_1) \tag{5.10}$$

where q_1 and q_3 are the lower and upper hinges (or quartiles), respectively. Recall that $\mathcal{E}_{th}$ and $\mathcal{E}_M$ are calculated only on the N_I samples corresponding to the final iteration, N_T, after the segmentation process has been completed.

5.3.4 Total probability formulation

With the total probability theorem, the probability of a shape to be reliable, $P(R = 1)$, can be estimated from the conditional probabilities with respect to r and its complementary, $\bar{r} = 1 - r$. In this way, L estimators are available for $P(R = 1)$, one per landmark. By averaging all of them we have:

$$\begin{aligned}\hat{P}\big(R(\hat{\mathbf{u}}_i) = 1\big) = \frac{1}{L}\Bigg(&\sum_{j=1}^{L} P\big(\hat{r}_i^{(j)}\big) P\big(R(\hat{\mathbf{u}}_i) = 1 \,|\hat{r}_i^{(j)}\big) + \\ &+ \sum_{j=1}^{L} P\big(\hat{\bar{r}}_i^{(j)}\big) P\big(R(\hat{\mathbf{u}}_i) = 1 \,|\hat{\bar{r}}_i^{(j)}\big)\Bigg)\end{aligned} \tag{5.11}$$

where $P\big(R(\hat{\mathbf{u}}_i)|\hat{r}_i^{(j)}\big)$ is the probability of the shape $\hat{\mathbf{u}}_i$ to be accurately segmented given that the *j-th* landmark is estimated as reliable by the appearance model.

The estimation of $P(R)$ can be performed only after appropriate training. This training involves two steps: 1) Estimating the involved probabilities, and 2) Determining the thresholds to map $\hat{P}(R=1)$ into the categories defined in the previous section, namely `reliable`, `unreliable` and `undefined`.

5.3.5 Conditional probabilities

The conditional probabilities in (5.11) measure how much can be known about the segmentation accuracy based on the estimated reliability of each landmark.

The estimation of the conditional probabilities can be computed from the training shapes $\mathbf{u_i}$ and their respective model fittings $\hat{\mathbf{u}}_i$ *at each iteration*. That is, the number of available samples per landmark is $N_I \times N_T$, being N_I the number of training images and N_T the number of model iterations. For each of these samples there is information about how precisely each landmark was located.

In most ASM-based approaches, the image search process is performed independently for each landmark. Indeed, it was pointed out in Section 5.2 that each landmark has its own image model. The coherence or *plausibility* of the shape is enforced later, at the regularization step (also known as shape restriction step) based on the Point Distribution Model (PDM).

Therefore, as long as the segmentation error in (5.3) is evaluated *before* the PDM restrictions are applied to the landmarks, it is reasonable to assume that the value of $\hat{r}_i^{(j)}$ computed by the appearance model of the *j-th* landmark is not affected by the position of any other landmark, but the *j-th*. In such a case, an estimator for $P(R=1)$ is derived in Appendix 5.7:

$$\hat{P}\big(R(\hat{\mathbf{u}}_i)=1\big) \simeq \frac{\sum_{j=1}^{L}\big(\hat{r}_i^{(j)}\rho^{(j|j)} + \hat{\bar{r}}_i^{(j)}\rho^{(j|\bar{j})}\big)\prod_{k\neq j}\rho^{(k)}}{\sum_{j=1}^{L}\prod_{k\neq j}\rho^{(k)}} \tag{5.12}$$

where $\rho^{(\cdot)}$ are ratios learnt over the training set for landmark j (see (5.33), (5.34) and (5.37) in Appendix 5.7). Specifically, if N_R were the samples estimated to be reliable, N_A the samples actually reliable (accurate), and N_{AR} the ones both accurate and estimated as reliable, then:

$$\rho^{(j)} = \frac{N_A}{N_T \times N_I} \tag{5.13}$$

$$\rho^{(j|j)} = \frac{N_{AR}}{N_R} \tag{5.14}$$

$$\rho^{(j|\bar{j})} = \frac{N_A - N_{AR}}{N_T \times N_I - N_R} \tag{5.15}$$

The ratios $\rho^{(j)}$ in (5.12) denote the *a priori* probability for a given sample to have the landmark j accurately placed. In the absence of any evidence, this probability

could just be assumed constant for all landmarks (i.e. $\frac{1}{2}$), transforming (5.12) into:

$$\frac{1}{L}\sum_{j=1}^{L}\left(\hat{r}_i^{(j)}\rho^{(j|j)} + \hat{\bar{r}}_i^{(j)}\rho^{(j|\bar{j})}\right) \tag{5.16}$$

The convenience of such simplification may be arguable, but it is useful to show that the result obtained for $\hat{P}(R=1)$ can also be interpreted as a weighted sum over the landmarks regarded as reliable based on the appearance model.. The weights, $\rho^{(j|j)}$ and $\rho^{j|(\bar{j})}$, account for the degree of correlation observed on the training set between the fitting error, $\epsilon_i^{(j)}$, and the estimated reliability of each landmark, $\hat{r}_i^{(j)}$.

This point of view brings up an interesting aspect regarding the two terms of the summations in (5.16). The first one, involving $\rho^{(j|j)}$, increases $\hat{P}(R=1)$ each time a landmark is estimated as reliable by the appearance model. Its weighting factors $\rho^{(j|j)}$ were estimated as the fraction of reliable cases ($\hat{r}_i^{(j)}=1$) for which the landmark was actually correctly placed ($r_i^{(j)}=1$).

Furthermore, the second term increases $\hat{P}(R=1)$ each time a landmark is declared non-reliable by the appearance model. Its weighting factors $\rho^{(j|\bar{j})}$ were estimated as the fraction of non-reliable cases ($\hat{r}_i^{(j)}=0$) for which, however, the landmark was correctly placed ($r_i^{(j)}=1$).

5.3.6 Defining reliability thresholds

Fig. 5.3 shows an example of estimated reliability versus segmentation error. Each point on the plot corresponds to a pair $\{\hat{P}(R=1);\mathcal{E}\}$ for one of the $N_T \times N_I$ samples mentioned before. The linear relationship between $\hat{P}(R=1)$ and $\mathcal{E}$ is shown by displaying the first Principal Component of the cloud of points (dashed line).

The horizontal line at $\mathcal{E}_{th}$ is the expected precision of the algorithm. Shapes whose segmentation error is smaller than $\mathcal{E}_{th}$ have completed a successful (or accurate) process. On the other hand, when the error is considerably higher than $\mathcal{E}_{th}$ the segmentation result is not accurate enough, which is indicated by $\mathcal{E}_M$.

From the linear relationship of $\hat{P}(R=1)$ and $\mathcal{E}$, the thresholds τ_1 and τ_2 get univocally determined once $\mathcal{E}_{th}$ and $\mathcal{E}_M$ have been set. The point colors on Fig. 5.3 indicate this latter division. The light-gray samples are estimated as `unreliable` and they are expected to be above $\mathcal{E}_M$. The darkest color indicates samples estimated as `reliable`, expected to be below $\mathcal{E}_{th}$. The remaining samples are `undefined`, possibly in the middle of a converging (or diverging) matching process. With the help of τ_1 and τ_2, $\hat{P}(R=1)$ can be mapped into these three classes, similarly to the division defined for R in Section 5.3.3. Therefore, such division on the horizontal axis provides an estimation to R, say $\hat{R}$, and $\hat{P}(R)=P(\hat{R})$.

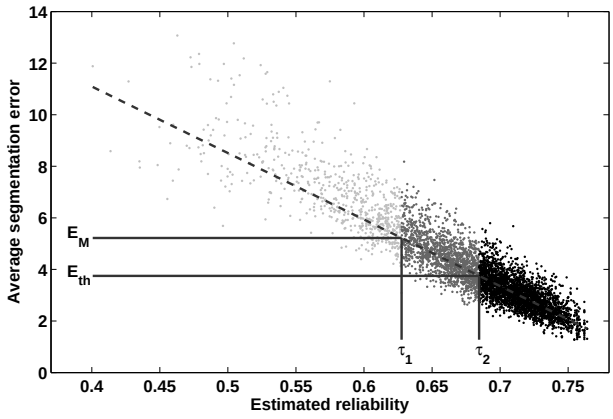

Figure 5.3: Example of reliability estimates and their respective segmentation errors, for segmentations of the XM2VTS database downsampled by a factor of 16. The different colors indicate how each sample is classified: `reliable` (black), `unreliable` (light-gray) or `undefined` (dark-gray).

During the segmentation of a new image not in the training set, the only information available will be the one on the horizontal axis. Thus, by computing $\hat{R}$ we will try to infer the accuracy of the automatic segmentation of this new image.

5.3.7 Incremental accumulation of reliability evidence

The formulation presented in the previous sections was concerned with the reliability of a certain shape fitted to an image (a single iteration in ASM-based approaches). However, we are interested in the final result of the segmentation process, which involves a number of iterations.

The iterative nature of ASMs can be useful in two aspects. Firstly, it can add a prior to the estimation of $P(R)$, as each iteration is clearly dependent on the previous one. Secondly, it enables for the *accumulation of evidence*, for example by using Bayesian chaining [8]. By adding the argument t to indicate the iteration number (as introduced in Section 5.3.1), Bayesian chaining can be formulated as:

$$P\big(R(\hat{\mathbf{u}}_i;t)\big) \simeq P\big(R(\hat{\mathbf{u}}_i;t-1)\big)\frac{P\big(\hat{R}(\hat{\mathbf{u}}_i;t)\,|\,R(\hat{\mathbf{u}}_i;t)\big)}{P\big(\hat{R}(\hat{\mathbf{u}}_i;t)\big)} \tag{5.17}$$

where $P\big(R(\hat{\mathbf{u}}_i;t)\big)$ – or simply $P\big(R(t)\big)$ – is the updated probability of accuracy of the segmentation given the new evidence provided by the estimation $P\big(\hat{R}(\hat{\mathbf{u}}_i;t)\big)$ available at iteration t. In terms of likelihoods, a proportionality relation can be

derived from (5.17):

$$P\big(R(t)\big) \sim \mathcal{L}\big(R(t) \,|\, \hat{R}(t)\big) \tag{5.18}$$

This formulation, however, must be corrected, as it does not account for the dependence between subsequent iterations. Therefore, in our case:

$$P\big(R(t)\big) \sim \mathcal{L}\big(\{R(t) \,|\, \hat{R}(t-1)\} \,|\, \{\hat{R}(t) \,|\, \hat{R}(t-1)\}\big) \tag{5.19}$$

The accumulation of evidence through the complete iterative process can then be summarized as a likelihood product. Thus, $P(R)$ is proportional to:

$$\mathcal{L}\big(R(1) \,|\, \hat{R}(1)\big) \prod_{t>1} \mathcal{L}\big(\{R(t) \,|\, \hat{R}(t-1)\} \,|\, \{\hat{R}(t) \,|\, \hat{R}(t-1)\}\big) \tag{5.20}$$

As it was already mentioned, there are three hypothesis available for R: `reliable`, `undefined` and `unreliable`. The one with the highest likelihood product is the class assigned to the whole iterative process.

5.3.8 Summary of the method

As it was previously stated, the objective of our reliability estimator is to determine whether a certain precess of model matching can be trusted or not. Algorithm 2 details the reliability estimation step by step.

Algorithm 2 Estimation of the reliability of the model-to-image matching

1: **for** (all iterations) $t = 1$ to N_T **do**
2: **for** (all landmarks) $j = 1$ to L **do**
3: Estimate $\hat{r}^{(j)}(t)$ from (5.8)
4: **end for**
5: Compute $\hat{P}\big(R(t)\big)$ from (5.12)
6: Compute (and store) $\hat{R}(t)$ by thresholding $\hat{P}\big(R(t)\big)$
7: **end for**
8: **for** (all hypotheses) $R = -1$ to 1 **do**
9: Initialize likelihood product to $\mathcal{L}\big(R(1) \,|\, \hat{R}(1)\big)$
10: **for** (all iterations) $t = 2$ to N_T **do**
11: Update likelihood product using (5.20)
12: **end for**
13: **end for**
14: Decide for the hypothesis with highest likelihood product

The proposed method starts off by estimating whether each landmark placement is individually considered reliable. Such decision is made based on the dissimilarity metric of the appearance models (lines 2 to 4 of Algorithm 2). An indicative example

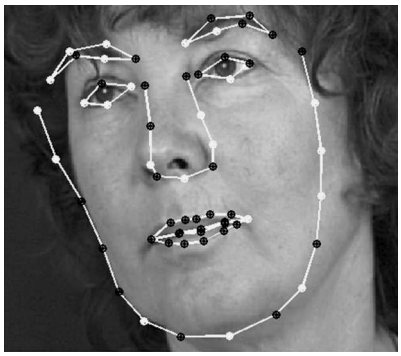

Figure 5.4: Example of the segmentation of a facial image from the XM2VTS database. The landmarks are displayed in different colors according to their estimated reliability (● = reliable, ○ = unreliable).

is presented in Fig. 5.4. It shows a model fitting to a facial image, where landmarks have been painted in different color depending on whether they were estimated as reliable (dark-gray) or not (light-gray).

At a second step, the whole shape is analyzed (line 5). It would be expected that in an *accurate* segmentation most landmarks would be reliable and vice-versa. Hence, the global reliability, R, of the shape is estimated as a weighted average of the individual reliability of all landmarks. The weights take into account how accurate individual estimates are. For example, in Fig. 5.4 the landmarks on the eyes seem to be accurately placed ($r = 1$), but some of them have been estimated as unreliable ($\hat{r} = 0$). On the other hand, the reliability estimations of the landmarks in the silhouette contour correlate better with the accuracy of their placement. If this behavior is consistent through the training set, then the estimations of the silhouette will have higher weights than those of the eyes.

The global reliability of the shape can be estimated at each iteration of the model fitting. This not only increases the number of samples available for the estimation, but also allows for monitoring the *behavior* of the segmentation process (lines 8 to 13).

The concept is illustrated with an example in Fig. 5.5. The upper plot shows the evolution of the segmentation error $\mathcal{E}(\hat{\mathbf{u}}_i; t)$ together with a marker indicating the reliability estimated at each iteration, $\hat{R}(t)$. These individual estimations are combined as indicated in (5.20) to calculate the likelihood products for the `reliable` and `unreliable` hypotheses, as showed in the lower plot.

At the beginning of the process, the segmentation error is high, and the estimated reliability low, enforcing the hypothesis that the segmentation is `unreliable`. As the model iterates, the error diminishes and the samples get more frequently regarded reliable by our method. Above iteration number 10 there is a long sequence

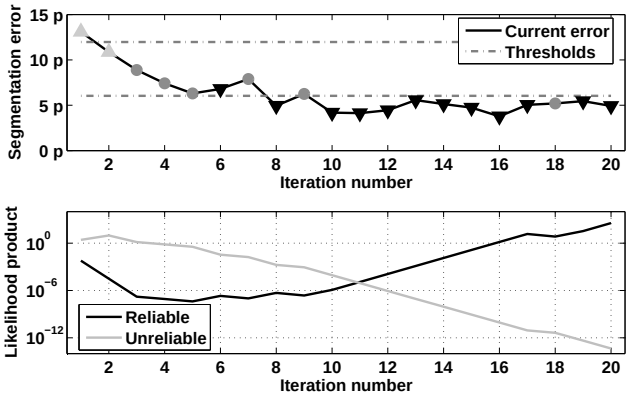

Figure 5.5: Example of the iterative segmentation process. The top plot shows the segmentation error (solid line) and the thresholds $\mathcal{E}_{th}$ and $\mathcal{E}_M$. The markers indicate whether each iteration was regarded as `reliable` (▼), `unreliable` (▲) or `undefined` (●). The bottom plot shows the evolution of the likelihoods for the whole process to be `reliable` (dark line) or `unreliable` (light line).

of `reliable` samples, strongly supporting the hypothesis of a `reliable` segmentation.

It is also interesting to analyze the precision of the global reliability estimates at each iteration. In the example of Fig. 5.5 most of the times the estimates are correct, but a few mistakes are also present (that is $\hat{R}(t) \neq R(t)$). For example, at iteration 2 the segmentation error was `undefined` while the sample was regarded as `unreliable`, and at iteration 6 the segmentation error was again `undefined` and this time it was estimated as `reliable`. These mistakes are not surprising. Indeed, they match what was shown in Fig. 5.3: there is a strong correlation between segmentation error and reliability estimates, but the prediction of $\mathcal{E}$ based on $\hat{P}(R)$ is not perfect.

The accumulation of evidence by means of (5.20) improves the robustness of the method against failures of the individual estimates. For example, in Fig. 5.5, the 18-*th* iteration was regarded as `undefined`, which may lead to an incorrect decision, with strong dependency on the last iteration analyzed. However, previous evidence filters the contribution of this sample, and suggests that the accumulated likelihood supporting the process being `reliable` is clearly higher than the one of being `unreliable`.

5.4 Segmentation results

In this section, experiments demonstrating the performance of the proposed method are presented. Three datasets were used: a subset of 532 images from the AR database [126], showing four different facial expressions for 133 individuals; a subset of 546 images from the Equinox database [171], containing 6 different images for 91 subjects, and the XM2VTS database [133], composed of 2360 images (8 for each of 295 individuals). The first two datasets were manually landmarked with a 98-point template [177]. For the XM2VTS database, annotations with a 64-point template were obtained from [1].

Three ASM and IOF-ASM models were constructed (one for each database). The parameters used for the construction of the models were the same detailed in [179], and the thresholds and coefficients for the reliability computation were estimated following Section 5.3. The training sets consisted in 180 images for the AR database, 186 for Equinox (as in [179]) and 400 images for the XM2VTS database, corresponding to the intersection of the training sets for both configurations of the Lausanne protocol [133].

The first experiment consisted in segmenting the images of all datasets and computing the corresponding reliability scores. Tables 5.1, 5.2 and 5.3 show the averaged point-to-curve distance (in pixels) for all the shapes (E_{ALL}), and separately for the ones regarded `reliable` (E_{rel}) and `unreliable` (E_{unrel}). The standard error is also indicated, as well as the number of images belonging to each class. The errors for each face were normalized dividing by 1% of the distance between the centers of the eyes (from the manual annotations). All segmentation errors reported in this paper were normalized following this criteria, to enable comparison across databases.

The correlation between the estimated reliability and the segmentation error is evident. In all cases there was a statistically significant difference between E_{rel} and E_{unrel} (a t-test showed confidence larger than 95% in all cases). The results are further illustrated in Fig. 5.6, focusing on the XM2VTS database. Notice the evident separation in the segmentation accuracy of the samples estimated as reliable and unreliable.

The curves in Fig. 5.6 are analogous to the error rate curves [21] used in verification experiments. In this case, the classes to separate are the accurate and inaccurate segmentations, and the threshold varying through the horizontal axis is the segmentation error. The black curves are the equivalent to the False Acceptance Rate (FAR), as they show the proportion of segmentations that have been considered `reliable` although their segmentation error exceeds the value indicated by the horizontal axis. And the light-gray curves show the False Rejection Rate (FRR): the proportion of segmentations misclassified as `unreliable`, since their segmentation error was below the value indicated by the horizontal axis. It is remarkable the very low FAR that can be obtained setting this threshold at about 6 pixels.

Table 5.1: Segmentation Errors [pixels] on AR Database

Model	Training Set	E_{ALL}	E_{rel}	E_{unrel}
IOF	**AR** (180 img)	1.63 ± 0.03 (532 img)	1.54 ± 0.03 (510 img)	3.52 ± 0.50 (22 img)
	Equinox (186 img)	4.37 ± 0.12 (532 img)	3.16 ± 0.11 (137 img)	4.79 ± 0.14 (395 img)
	XM2VTS (400 img)	4.21 ± 0.12 (532 img)	2.14 ± 0.09 (97 img)	4.67 ± 0.14 (435 img)
ASM	**AR** (180 img)	2.42 ± 0.06 (532 img)	2.16 ± 0.04 (505 img)	6.09 ± 0.67 (27 img)
	Equinox (186 img)	4.64 ± 0.09 (532 img)	2.84 ± 0.13 (49 img)	4.82 ± 0.10 (483 img)
	XM2VTS (400 img)	5.17 ± 0.14 (532 img)	4.02 ± 0.11 (372 img)	7.77 ± 0.28 (160 img)

A careful analysis of Tables 5.1, 5.2 and 5.3 suggests another important conclusion: the accuracy of the `reliable` segmentations was better extrapolated to a different database than the overall segmentation error. For example, the IOF-ASM model trained with the AR database achieved an average segmentation error of 1.63 pixels on the AR database, which increased to 5.22 pixels on the XM2VTS database (+220%). On the other hand, the error for the segmentations estimated as `reliable` increased only from 1.54 to 2.75 pixels. The same trend was observed for both models in the three datasets employed.

5.4.1 Implementational issues

In this section we describe a number of details that, for the sake of clarity, were omitted from the explanation of the method as they mostly regard to its implementation.

Computational complexity

The computational load of the method is clearly concentrated at the training stage. Specifically, the computation of (5.3) and (5.4) involves running the automatic segmentation algorithm on all training images and computing the chosen distance metric with respect to the ground-truth at each iteration. The contribution of the remaining steps is negligible, including the maximization of (5.9), even if carried out by brute force search.

During segmentation, the reliability estimation only requires computing $\hat{P}(R)$ as

Table 5.2: Segmentation Errors [pixels] on Equinox Database

Model	Training Set	E_{ALL}	E_{rel}	E_{unrel}
IOF	AR (180 img)	3.59 ± 0.08 (546 img)	2.89 ± 0.06 (235 img)	4.12 ± 0.12 (311 img)
	Equinox (186 img)	1.92 ± 0.03 (546 img)	1.90 ± 0.02 (536 img)	2.83 ± 0.28 (10 img)
	XM2VTS (400 img)	4.08 ± 0.11 (546 img)	3.05 ± 0.09 (110 img)	4.34 ± 0.13 (436 img)
ASM	AR (180 img)	3.64 ± 0.08 (546 img)	3.60 ± 0.08 (541 img)	7.17 ± 1.04 (5 img)
	Equinox (186 img)	2.28 ± 0.05 (546 img)	2.19 ± 0.04 (525 img)	4.38 ± 0.34 (21 img)
	XM2VTS (400 img)	4.39 ± 0.11 (546 img)	4.29 ± 0.10 (533 img)	8.57 ± 1.68 (13 img)

in (5.12) for each iteration, and the likelihood products by means of (5.20). The computational time involved is therefore considerably smaller than the time required for model fitting of any ASM-like approach.

Multiple resolutions

In statistical shape models it is common to use a course-to-fine search strategy in a multi-resolution approach [45]. It consists on building a multi-resolution pyramid from the original image, and processing each level independently. Therefore, (5.20) can be computed separately at each resolution level, and the results combined together for the final decision of the whole segmentation process. For the experiments reported in this paper we simply averaged all N_S resolution levels:

$$\hat{R}^{(final)}(\hat{\mathbf{u}}) = \frac{1}{N_S} \sum_{k=1}^{N_S} \hat{R}^{(k)}(\hat{\mathbf{u}}) \tag{5.21}$$

where $\hat{R}^{(k)}$ takes values within the set $\{-1, 0, 1\}$, according to the winning hypothesis for the *k-th* resolution level (the one with the highest likelihood in (5.20): "1" = `reliable`, "0" = `undefined`, "-1" = `unreliable`). The segmentation process is then regarded as `reliable` if $\hat{R}^{(final)}(\hat{\mathbf{u}})$ is positive, and `unreliable` otherwise.

Notice that this combination rule is the simplest possible, and does not take into account that each resolution level is likely to have a different precision in the estimation of reliability. Additionally, finer resolutions are likely to have lower values of $\mathcal{E}_{th}$, which would be a further hint for combination. On the other hand, more complicated rules based on the training set would suffer from little data samples.

Table 5.3: Segmentation Errors [pixels] on XM2VTS Database

Model	Training Set	E_{ALL}	E_{rel}	E_{unrel}
IOF	AR (180 img)	5.22 ± 0.10 (2360 img)	2.75 ± 0.04 (919 img)	6.53 ± 0.16 (1441 img)
	Equinox (186 img)	9.52 ± 0.19 (2360 img)	3.66 ± 0.08 (534 img)	11.24 ± 0.23 (1826 img)
	XM2VTS (400 img)	2.03 ± 0.02 (2360 img)	1.92 ± 0.01 (2214 img)	3.67 ± 0.18 (146 img)
ASM	AR (180 img)	9.62 ± 0.20 (2360 img)	3.81 ± 0.10 (513 img)	11.1 ± 0.24 (1847 img)
	Equinox (186 img)	13.9 ± 0.29 (2360 img)	2.06 (1 img)	13.9 ± 0.29 (2359 img)
	XM2VTS (400 img)	3.06 ± 0.08 (2360 img)	2.69 ± 0.04 (2141 img)	6.80 ± 0.72 (219 img)

Indeed, the N_T iterations are already combined in (5.20), and only N_I samples remain available. For example, in the experiments of the next section, N_I will be as small as 30.

Negative correlation

An underlying assumption that must be verified when building a reliability model is the negative correlation between $\mathcal{E}(\hat{\mathbf{u}})$ and $\hat{P}(R)$, so that lower segmentation errors are expected to produce higher reliability estimates. Would this not be true for a given resolution level, then that level is useless to estimate reliability. For example, the average correlation between $\mathcal{E}(\hat{\mathbf{u}})$ and $\hat{P}(R)$ on the training set of the XM2VTS database was of -0.75 for IOF-ASM and -0.61 for ASM, clearly fulfilling the above requirement.

Likelihood normalization

When evaluating (5.20), there are three possible hypothesis for $R(\hat{\mathbf{u}})$: `reliable`, `unreliable` and `undefined`. Notice that the actual value of the likelihoods is not important as long as they preserve the relationship between them. Therefore, it is possible to *normalize* the likelihoods with respect to any of the hypothesis. Let $\mathcal{L}_N(t,h)$ be the normalized likelihood for hypothesis $R(\hat{\mathbf{u}};t) = h$, and $h = 0$ be the

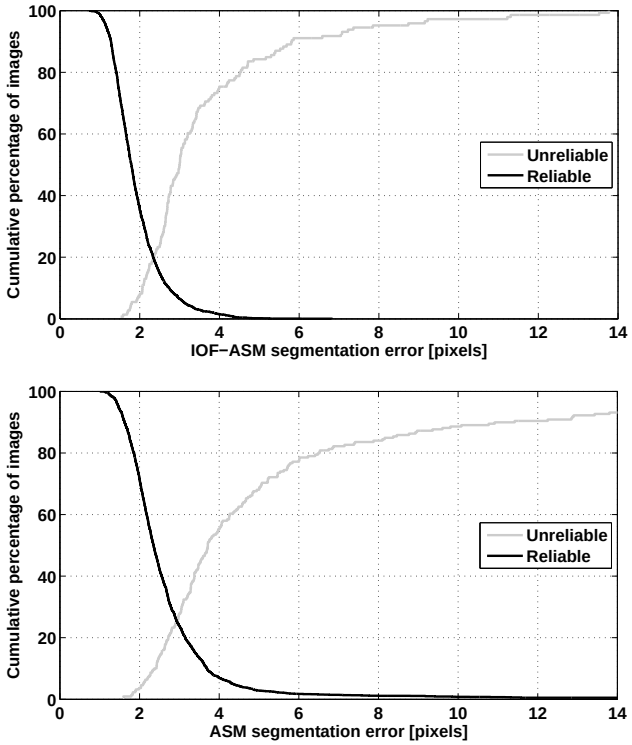

Figure 5.6: Segmentation error statistics for the XM2VTS database. The light-gray curves show the percentage of images estimated as unreliable whose error was less than (or equal to) the value indicated by the horizontal axis. The black curves show the percentage of images estimated as reliable whose error was greater than the value of the horizontal axis.

normalizing hypothesis. Then we write:

$$\mathcal{L}_N(t,h) = \frac{\mathcal{L}\big(\{R(t)=h \,|\, \hat{R}(t-1)\} \,|\, \{\hat{R}(t) \,|\, \hat{R}(t-1)\}\big)}{\mathcal{L}\big(\{R(t)=0 \,|\, \hat{R}(t-1)\} \,|\, \{\hat{R}(t) \,|\, \hat{R}(t-1)\}\big)} \tag{5.22}$$

$$P(R) \sim \frac{\mathcal{L}\big(R(1) \,|\, \hat{R}(1)\big)}{\mathcal{L}\big(R(1)=0 \,|\, \hat{R}(1)\big)} \prod_{t>1} \mathcal{L}_N(t,h) \tag{5.23}$$

The likelihoods displayed in the lower plot of Fig. 5.5 have been normalized in this way, so the curve corresponding to `undefined` ($R = 0$) is just a constant at 10^0. This implies that, from iterations 4 to 16, the winning hypothesis in Fig. 5.5 would be $R = 0$.

However, the upper plot of Fig. 5.5 suggests that both the segmentation error

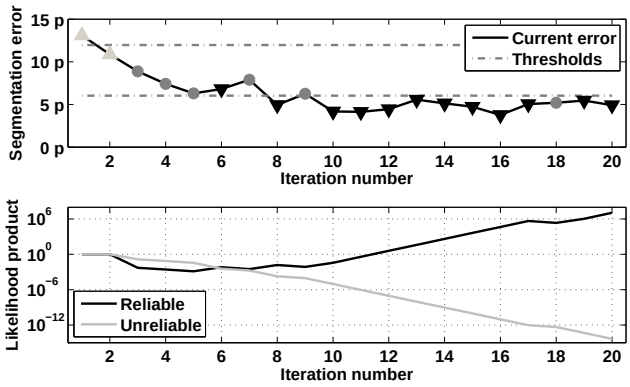

Figure 5.7: Example of the iterative segmentation process. The upper plot shows the segmentation error (solid line) and the thresholds $\mathcal{E}_{th}$ and $\mathcal{E}_M$. The markers indicate whether each iteration was regarded as `reliable` (▼), `unreliable` (▲) or `undefined` (•). The lower plot shows the evolution of the likelihood for the whole process to be `reliable` (dark line) or `unreliable` (light line). Likelihoods are taken into account starting at the third iteration ($t_0 = 3$, see text).

and the global reliability estimation have stabilized (or converged) much earlier. This *delay* in deciding toward the `reliable` class obeys to the bias introduced by the first few iterations: in general, the segmentation error at the beginning of the process is higher than the required accuracy, which favors the likelihood of the `unreliable` class. The bias of the first iterations can be reduced by slightly modifying (5.23):

$$P\{\mathbf{R}\} \sim \prod_{t \geq t_0} \mathcal{L}_N(t, h) \tag{5.24}$$

where t_0 is the first iteration at which the hypothesis of `unreliable` segmentation does not have the highest likelihood:

$$\mathcal{L}_N(t_0, -1) < 1 \tag{5.25}$$

$$\mathcal{L}_N(t, -1) \geq 1 \qquad \forall t < t_0 \tag{5.26}$$

All the results reported in this paper are based on (5.24). Note that, if the number of iterations is large enough, then (5.23) and (5.24) produce the same result, but the latter is expected to *converge* faster. This is shown in Fig. 5.7 where the example of Fig. 5.5 is repeated using (5.24). As the first two iterations do not bias the evaluation of likelihoods, the hypothesis of an `accurate` segmentation is achieved 5 iterations earlier than in the case of Fig. 5.5, computed by means of (5.23).

Figure 5.8: Sample images of one individual from the AV@CAR dataset. From left to right: R60-, R40-, R20- and Frontal-views (top row); L60-, L40- and L20-views (bottom row).

5.5 Applications

5.5.1 Automatic model selection

A straightforward application of a reliability measure is to automatically select, from a library of models, the best model to segment a given image. Let $\{\mathcal{M}_k\}$ be a set of statistical shape models and $\{I_i,\}$ a set of images to be segmented. For each image I_i there is a ground-truth shape $\mathbf{u}_i$. We want to determine which is the best model to use for each of the images:

$$k_i^{(best)} = \underset{k}{\operatorname{argmin}} \left(d\big(\mathbf{u}_i, \mathcal{M}_k\{I_i\}\big) \right) \tag{5.27}$$

where $k_i^{(best)}$ is the index of the best model and $\mathcal{M}_k\{I_i\}$ is the shape generated by the model $\mathcal{M}_k$ when segmenting image I_i. Evidently, the objective is to estimate $k_i^{(best)}$ as accurately as possible without knowing $\mathbf{u}_i$, which can be done in a direct way by choosing the model whose segmentation reported the highest reliability.

To demonstrate the above statement, a set of 280 images from the AV@CAR database [143] was used, showing 7 views for each of 40 subjects. The views were systematically selected to vary according to the degree of lateral head rotation (out of the image plane). The resulting set is illustrated in Fig. 5.8 and contains three rotated views to each side (L20, L40, L60 and R20, R40, R60; at approximately ± 20, ± 40 and ± 60 degrees) and one *frontal* view (F), at which the subject is looking straight into the camera.

For each of the segmentation algorithms, ASM and IOF-ASM, we constructed seven models, one for each view. All images in the dataset were segmented with each of these single-view models, computing also the reliability scores in every case. Then, the most reliable result was chosen. As there were only 40 images available

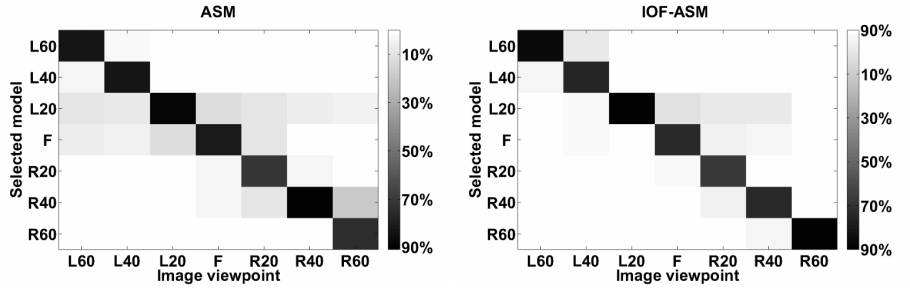

Figure 5.9: Confusion matrices for ASM and IOF-ASM in the automatic model selection tests. The horizontal axis shows the viewpoint of the test images, while the vertical axis shows the different models. At each position of the grid, the color code indicates the percentage of images that were automatically assigned each of the models, following the reliability criterion (see text),

per view, a 4-fold cross validation was performed to split them into training and test sets. Thus, the models for each view were always built in sets of 30 images, disjoint from the set of test images.

Fig. 5.9 shows the confusion matrix between the view of each image and the model that produced the most reliable score. As it was expected, in most cases the selected model matched the view of the test image. Analyzing the results as a viewpoint recognition test, the achieved accuracy was of 89.6% for IOF-ASM and 82.1% for ASM.

Table 5.4 provides the segmentation results using four different strategies:

* The first row provides the best results achieved using just one of the single-view models (always the same) to segment all images. Both for ASM and IOF-ASM, the frontal-view model was the one that achieved the lowest segmentation error, as can be intuitively expected due to the symmetry of the data set.

* For the second row, we also used only one model. However, in this case it was trained with images from all the views. Therefore, $30 \times 70 = 210$ images were used for training (always under a 4-fold cross validation).

* In the third row, for each image the model that showed the most reliable result was selected. This is the strategy proposed in the above paragraphs of this section.

* Finally, we also report the results that could be obtained if the automatic selection would perfectly match the viewpoint of the test image (fourth row).

The analysis of the data in Table 5.4 clearly shows that the automatic selection can be very useful to improve segmentation accuracy. The large difference with

Table 5.4: Segmentation errors (average $\pm$ std. error, in pixels) on the AV@CAR dataset with different segmentation strategies

Segmentation strategy	ASM	IOF-ASM
Best single-view model	9.90 ± 0.46	7.47 ± 0.42
Multi-view model	7.26 ± 0.35	2.10 ± 0.07
Automatic model selection	2.79 ± 0.19	1.99 ± 0.11
Optimal model selection	2.10 ± 0.07	1.67 ± 0.03

respect to the best single-view model further indicates that an incorrect selection of the model will dramatically drop segmentation performance.

It is also interesting to compare the second and third rows of the table. In the case of ASM, the automatic selection largely outperforms the multi-view model. Indeed, other researchers have already reported that ASM performs well only within $\pm$20 degrees of left-right head rotation [112, 158]. This has led to split of the head rotation range into a number of narrower bands, in a strategy analog to the one proposed here. Those approaches, however, either did not address automatic model selection [203], or were based on AAMs and simply used the convergence of the represented texture as the selection criteria [47, 112].

In the case of IOF-ASM, the difference between the second and third rows of Table 5.4 is very small. This suggests that the more complex image model of IOF-ASM may be able to deal with a wider range of head rotation than ASM. Nonetheless, automatic model selection allows for extending the rotation range to poses at which not all landmarks are visible. To achieve this with just one multi-view model there would be the need to modify the PDM, for example by resorting to non-linear statistics [158].

5.5.2 Reliable identity verification

Reliability estimation could be used to enhance the robustness of an identity verification system. A crucial problem in many fully automatic systems for identity verification is the occasional failure of some of their processing blocks. For example, Fig. 5.1 shows the segmentation results for three different initializations of a statistical shape model. Two of them are clearly of no use for a verification system. No matter whether the failure is due to an incorrect initialization or due to a failure within the segmentation process, automatically discriminating between reliable and unreliable results can be a valuable aid in increasing system robustness: if the segmentation of a given image is not reliable, then the system could request a new frame, or take the default (minimum risk) decision according to the specific application.

To verify the above hypothesis, identity verification tests were performed on the XM2VTS database, using both configurations of the Lausanne protocol [133]. The *evaluation* sets Equal Error Rates (EER) [21] were used to set the working point of the classifier, at which the Half Total Error Rates (HTER) were computed for the tests sets.

The biometric parameters for each image were obtained by PCA of the face texture (the image region enclosed by the landmarks found by the automatic segmentation), after a warping to the mean shape of the training set [106]. Then, a classifier based on whitened correlation was used, known to be among the best choices for PCA-based metrics [145].

Initialization

As it was pointed out in [176], the segmentation accuracy of statistical shape models on the XM2VTS database is enough to achieve results comparable to the ones using manual annotations. However, the situation changes under more challenging initializations, creating a more appropriate scenario for evaluating the proposed reliability measure.

With this idea in mind, segmentation tests were repeated on the XM2VTS database for 49 different initial positions, resulting from a uniform 7×7 grid. The central position of the grid matched the initialization used in Section 5.4, which was always the shape centroid computed from the manual annotations (we will refer to this position as the *optimal* location). Notice, however, that the initial shape was always the mean shape of the training set with a fixed size, therefore resulting in a different initialization error for each image (see also [179]). The remaining initializations were all possible combinations of up/down and left/right displacements of the centroid by 20, 40 and 60 pixels, as illustrated in Fig. 5.10.

Results

Before presenting the verification scores, it is interesting to look at the reliability results from the segmentations, shown in Fig. 5.11. It can be seen that, as the initialization departs from the optimal location, the number of unreliable segmentations tends to grow, getting close to 100% for the maximum displacements. The segmentation error (with respect to the ground-truth from manual annotations) exhibited a similar behavior, both for ASM and IOF-ASM algorithms. Additionally, the direction of the displacements did not seem to noticeably influence performance.

Fig. 5.12 presents the variation of the HTER for the different initializations using IOF-ASM segmentation. The displayed curves were constructed using configuration II of the Lausanne protocol. The behavior of both ASM and IOF-ASM under configuration I was completely analogous, with slightly higher error rates than in Configuration II. Three curves are displayed:

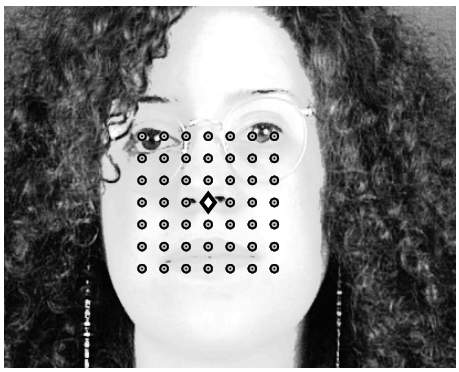

Figure 5.10: The 49 initial centroids for the segmentation process in the identity verification experiments. The diamond indicates the *optimal* initialization, corresponding to the centroid from the manual annotations. Both vertical and horizontal displacements are located in steps of 20 pixels, resulting in a maximum distance of 84.9 pixels with respect to the central diamond.

1. *All images*: HTER computed from the 1560 images, using automatic segmentation.

2. *Reliable images*: HTER computed only from the images whose segmentation was estimated to be `reliable`.

3. *Ground-truth*: HTER computed from the 1560 images using manual annotations.

The error bars indicate a 90% confidence interval using the *test of two proportions* [12], not shown on the ground-truth curve for clarity reasons. The 800 client images defined for *training* in the Lausanne protocol were assumed to be correctly segmented, as they would be the enrolled users on the verification database. Hence the manual annotations were used for those images in all experiments. The other 1560 images (evaluation and test sets for clients and impostors) were automatically segmented and, for the second case (*Reliable images*), discarded when estimated `unreliable`.

For initializations up to 40 pixels away from the optimal positioning, there is no statistically significant difference between the three curves. But for a displacement above 60 pixels, the scores using all images degrade rapidly, as opposed to the scores using only reliable segmentations, which remain almost unchanged. At the maximum displacement (when almost all images were estimated unreliable, see Fig. 5.11), the error rate approaches 50% (random choice).

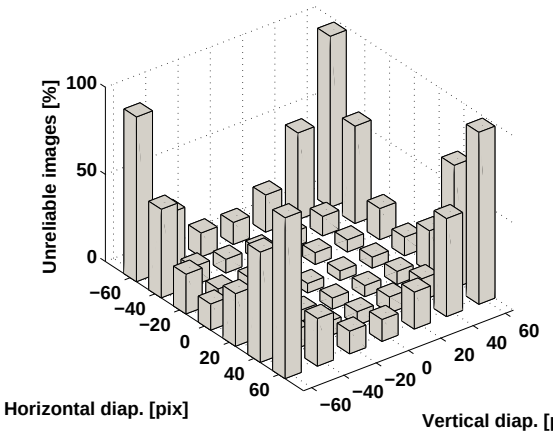

Figure 5.11: Percentage of `unreliable` segmentations using IOF-ASM for different initializations. The displacements are measured with respect to the *optimal* location, computed as the centroid of the manual annotations.

Keeping only the reliable segmentations yields significantly better verification scores. In fact, even for the largest initial displacement, the HTER obtained with the reliable images is not significantly different from the zero-displacement HTER, due to the very few samples available for its computation.

Fig. 5.13 provides the results obtained using ASM segmentation. In this case, the error rates of the reliable images show a significant degradation for displacements above 40 pixels, but they are still much better than those achieved using all the images.

Analyzing the segmentation errors of the experiments in this section provides further evidence to understand the reduction of the error rates when discarding non-reliable segmentations. Table 5.5 shows the errors averaged on the $115,640$ segmentations (49 initializations $\times$ 2360 images). When discriminated by their estimated reliability, differences are evident. Furthermore, it can be seen that the accuracy of the `reliable` samples was kept very similar to the one reported in Table 5.3 (which corresponds to optimal initialization). Fig. 5.14 illustrates typical segmentation results, corresponding to the average accuracies indicated in Table 5.5.

5.6 Summary and conclusions

A method to estimate the reliability of image segmentation with statistical shape models was presented. By means of a probabilistic framework, the information

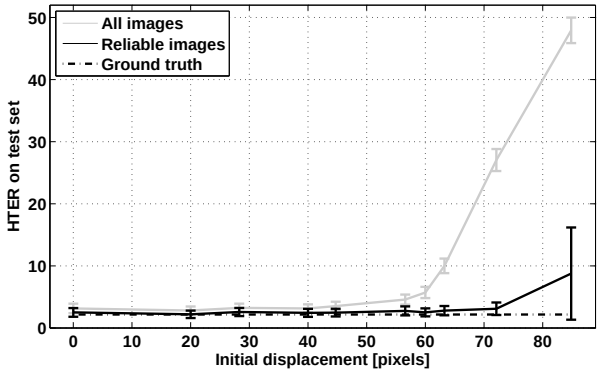

Figure 5.12: Half Total Error Rate using IOF-ASM segmentation for different initial positions. As the results did not reveal a major sensitivity to the direction of the displacements, the horizontal axis shows the distance with respect to the optimal location.

Table 5.5: Segmentation errors on the XM2VTS database with 49 different initialization displacements

Segmentation error computed on	ASM	IOF-ASM
Reliable segmentations	3.20 pix.	1.92 pix.
Unreliable segmentations	37.63 pix.	27.31 pix.
All available segmentations	24.83 pix.	8.11 pix.

provided by each landmark was combined to make a global decision regarding the quality of the whole model matching. The proposed method can be applied to statistical shape models in general, as long as a dissimilarity measure based on appearance is available for every landmark.

Segmentation results were provided for ASM and IOF-ASM using three different face databases. In all cases the estimated reliability exhibited a high correlation with the segmentation error. When extrapolating the models to a different database that the one used for training, the accuracy for the reliable matches was less degraded than the overall accuracy.

The automatic estimation of reliability can be promising for a number of applications. We demonstrated two of them: automatic model selection and reliable identity verification. Results were highly satisfactory in both cases.

The strength of the proposed approach relies on its low false positive rate. Indeed, the estimation of reliability does not increase the accuracy of the algorithms, but rather automatically assesses their performance. If the models perform poorly

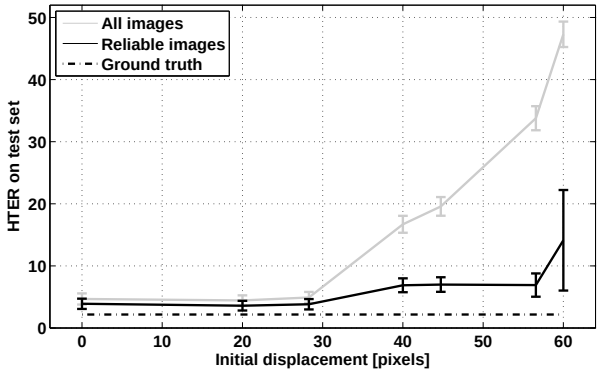

Figure 5.13: Half Total Error Rate using ASM segmentation for different initial positions. As the results did not reveal a major sensitivity to the axis or the direction of the displacements, the horizontal axis shows the distance with respect to the optimal position.

(like the segmentation of XM2VTS images with a model trained on Equinox images as shown in Table 5.3), then very few samples are declared reliable.

On the application side, this can be very useful. The initialization displacements shown in Fig. 5.10 are not dramatic (when compared to the size of the face) and however they dropped the performance of a typical identity verification system from error rates close to the state-of-the-art to error rates close to 50%. When the non-reliable segmentations were discarded, performance was kept considerably more stable.

It should be pointed out that this increased robustness of the system is not limited to an incorrect initialization. Any failure leading to a segmentation different enough from the training data would result in a non-reliable classification. That is the reason why, in the model selection experiments, the segmentation of images under a certain viewpoint only occasionally showed reliable results when segmented with models constructed from a different viewpoint. As shown in Fig. 5.9, the more separated the viewpoints, the less likely for this situation to happen.

5.7 Appendix: conditional probabilities

From the definitions in Sections 5.3.1 and 5.3.3, the reliability R is related to the averaged segmentation error for all landmarks:

$$P\big(R(\hat{\mathbf{u}};t)=1\big) \triangleq P\big(\mathcal{E}_i(\hat{\mathbf{u}}_\mathbf{i};t) < \mathcal{E}_{th}\big) = P\bigg(\sum_{j=1}^{L}\epsilon_i^{(j)}(t) < \sum_{j=1}^{L}\epsilon_{th}^{(j)}\bigg) \tag{5.28}$$

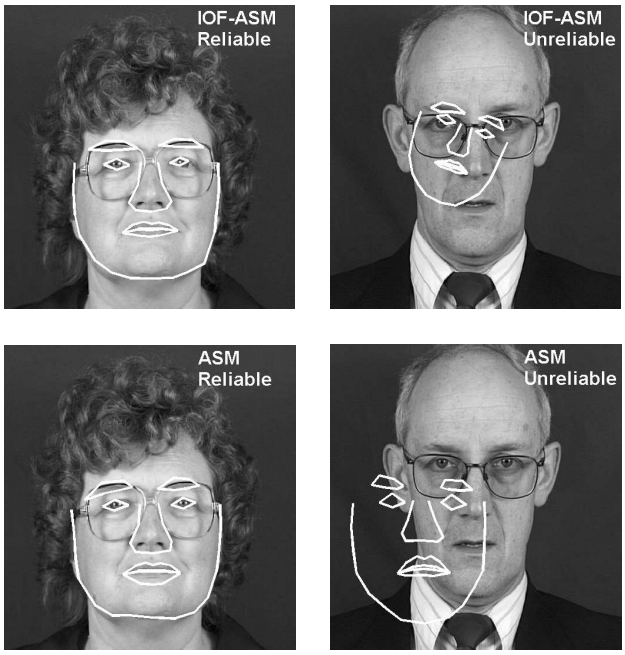

Figure 5.14: Typical results for the reliable and unreliable segmentations on the multiple-initializations experiments. The errors of the displayed images match the averages reported in Table 5.5 for each algorithm.

It can be made more restrictive by requiring the error *for each landmark* to be below a threshold. Let $R'(\hat{\mathbf{u}}_i)$ be this new measure, defined as:

$$\begin{aligned} P\big(R'(\hat{\mathbf{u}}_i;t)=1\big) &\triangleq P\big(\epsilon_i^{(j)}(t) < \epsilon_{th}^{(j)}\big), \quad \forall j \\ &= P\Big(\bigcap_{j=1}^{L} \{\epsilon_i^{(j)}(t) < \epsilon_{th}^{(j)}\}\Big) \end{aligned} \tag{5.29}$$

where the iteration number has been made explicit by using t, as explained in Section 5.3.1.

The conditional probabilities of $R'(\hat{\mathbf{u}}_i)$ given the reliability estimated by the appearance model of each landmark are then:

$$P\big(R(\hat{\mathbf{u}}_i;t)=1 \,|\, \hat{r}_i^{(j)}(t)\big) = P\Big(\Big[\bigcap_{k=1}^{L} \big(\epsilon_i^{(k)}(t) < \epsilon_{th}^{(k)}\big)\Big] \,|\, \hat{r}_i^{(j)}(t)\Big) \tag{5.30}$$

and replacing by the definition of $r_i^{(k)}(t)$ (see (5.5)):

$$P\big(R(\hat{\mathbf{u}}_i;t)=1\,|\,\hat{r}_i^{(j)}(t)\big)=P\Bigg(\Big[\bigcap_{k=1}^{L} r_i^{(k)}(t)\Big]\,|\,\hat{r}_i^{(j)}(t)\Bigg) \tag{5.31}$$

Assuming that the reliability of each landmark is independent of that of any other landmark and using the definition of conditional probabilities, we immediately get:

$$P\big(R'(\hat{\mathbf{u}}_i;t)=1\,|\,\hat{r}_i^{(j)}(t)\big)=P\big(r_i^{(j)}(t)\,|\,\hat{r}_i^{(j)}(t)\big)\prod_{k\neq j}P\big(r_i^{(k)}(t)\big) \tag{5.32}$$

The conditional probabilities for $k=j$ can be estimated from the training set, by defining:

$$\rho^{(j|j)}\triangleq\hat{P}\big(r_i^{(j)}(t)\,|\,\hat{r}_i^{(j)}(t)\big) \tag{5.33}$$

$$\rho^{(j|\bar{j})}\triangleq\hat{P}\big(r_i^{(j)}(t)\,|\,\hat{\bar{r}}_i^{(j)}(t)\big) \tag{5.34}$$

which can easily be computed due to the binary nature of the involved variables:

$$\rho^{(j|j)}\sum_{\forall i,t}\hat{r}_i^{(j)}(t)=\sum_{\forall i,t}\hat{r}_i^{(j)}(t)\,r_i^{(j)}(t) \tag{5.35}$$

$$\rho^{(j|\bar{j})}\sum_{\forall i,t}\hat{\bar{r}}_i^{(j)}(t)=\sum_{\forall i,t}\hat{\bar{r}}_i^{(j)}(t)\,r_i^{(j)}(t) \tag{5.36}$$

To take into account the possibility of a biassed training set, $P\big(r_i^{(k)}\big)$ can be also estimated:

$$\rho^{(k)}\triangleq\hat{P}\big(r_i^{(k)}(t)\big)\triangleq\frac{\sum_{\forall i,t}r_i^{(k)}(t)}{\sum_{\forall i,t}\hat{r}_i^{(k)}(t)+\sum_{\forall i,t}\hat{\bar{r}}_i^{(k)}(t)} \tag{5.37}$$

In spite of their different definition, both $R(\hat{\mathbf{u}}_i)$ and $R'(\hat{\mathbf{u}}_i)$ can be estimated with the expression in (5.11). Hence, from (5.32):

$$\begin{aligned}\hat{P}\big(R'(\hat{\mathbf{u}}_i;t)\big)=\frac{1}{L}\sum_{j=1}^{L}\Bigg(&P\big(\hat{r}_i^{(j)}(t)\big)P\big(r_i^{(j)}(t)\,|\,\hat{r}_i^{(j)}(t)\big)\prod_{k\neq j}P\big(r_i^{(k)}(t)\big)+\\&+P\big(\hat{\bar{r}}_i^{(j)}(t)\big)P\big(r_i^{(j)}(t)\,|\,\hat{\bar{r}}_i^{(j)}(t)\big)\prod_{k\neq j}P\big(r_i^{(k)}(t)\big)\Bigg)\end{aligned} \tag{5.38}$$

By including the estimations from the training set, we can write:

$$\hat{P}\big(R'(\hat{\mathbf{u}}_i;t)\big)=\frac{1}{L}\sum_{j=1}^{L}\big(\hat{r}_i^{(j)}(t)\,\rho^{(j|j)}+\hat{\bar{r}}_i^{(j)}(t)\,\rho^{(j|\bar{j})}\big)\prod_{k\neq j}\rho^{(k)} \tag{5.39}$$

From (5.28) and (5.29) it can be shown that $P(R') \leq P(R)$, so (5.39) can be thought as a conservative lower bound to $\hat{P}(R)$.

As it is shown in Section 5.3.6, the actual value of $\hat{P}\{R\}$ is not essential. There is the need, instead, for an estimator that is *proportional* to $P\{R\}$. Taking this into account, our estimator $\hat{P}\big(R(\hat{\mathbf{u}}_i; t)\big)$ can be computed by:

$$\frac{\sum_{j=1}^{L} \left(\hat{r}_i^{(j)}(t)\rho^{(j|j)} + \hat{\bar{r}}_i^{(j)}(t)\rho^{(j|\bar{j})}\right) \prod_{k\neq j} \rho^{(k)}}{\sum_{j=1}^{L} \prod_{k\neq j} \rho^{(k)}} \tag{5.40}$$

where the division by the summation over all $\rho^{(j)}$ was added to scale $\hat{P}(R)$ into the range $[0, 1]$.

Bibliography

[1] Manual annotation of the XM2VTS face images (publicly available at http://www.isbe.man.ac.uk/∼bim/data/xm2vts/xm2vts_markup.html).

[2] G.A. Abrantes and F. Pereira. MPEG-4 facial animation technology: survey, implementation, and results. *IEEE Transactions on Circuits and Systems For Video Technology*, 9(2):290–305, 1999.

[3] Y. Adini, Y. Moses, and S. Ullman. Face recognition: the problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[4] J.K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: a review. In *Proc. 3rd IEEE Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, USA*, pages 2–14, 1994.

[5] I. Aizenberg, N. Aizenberg, and J. Vandewalle. *Multi-valued and universal binary neurons: theory, learning, applications*. Kluwer Academic Publishers, Boston/Dordrecht/London, 2000.

[6] I. Aizenberg, C. Butakoff, V. Karnaukhov, N. Merzlyakov, and O. Milukova. Blurred image restoration using the type of blur and blur parameters identification on the neural network. In *Proc. SPIE Image Processing: Algorithms and Systems, San José, CA, USA*, volume 4667, pages 460–471, 2002.

[7] E. Alpaydin. Multiple neural networks and weighted voting. In *Proc. 11th Int. Conf. on Pattern Recognition, The Hague, The Netherlands*, volume 2, pages 29–32, 1992.

[8] T. Arbel and F.P. Ferrie. On the sequential accumulation of evidence. *International Journal of Computer Vision*, 43(3):205–230, 2001.

[9] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *International Jorunal of Computer Vision*, 45(6):891–923, 1998.

[10] G. Behiels, F. Maes, D. Vandermeulen, and P. Suetens. Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models. *Medical Image Analysis*, 6(1):47–62, 2002.

[11] P. Belhumeur and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1996.

[12] S. Bengio and J. Mariéthoz. A statistical significance test for person authentication. In *Proc. ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo, Spain*, pages 237–244, 2004.

[13] D. Beymer. Face recognition under varying pose. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Seattle, WA, USA*, pages 756–761, 1994.

[14] D. Beymer and T. Poggio. Face recognition from one example view. In *Proc. 5th IEEE Int. Conf. on Computer Vision, Boston, MA, USA*, volume 19, pages 500–507, 1995.

[15] J.A. Black, M. Gargesha, K. Kahol, P. Kuchi, and S. Panchanathan. A framework for performance evaluation of face recognition algorithms. In *Proc. SPIE Internet Multimedia Management Systems, Boston, MA, USA*, volume 4862, pages 163–174, 2002.

[16] M.J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proc. 5th IEEE Int. Conf. on Computer Vision, Boston, MA, USA*, pages 374–381, 1995.

[17] D.M. Blackburn, J.M. Bone, and P.J. Phillips. Face recognition vendor test 2000. Technical report, DoD Counterdrug Technology Development Program Office Defense Advanced Research Projects Agency, National Institute of Justice, USA, 2001.

[18] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. 26th ACM SIGGRAPH Int. Conf. on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA*, pages 187–194, 1999.

[19] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[20] W. Bledsoe. The model method in facial recognition. Technical report, 49 Panoramic Research Inc., Palo Alto, CA, USA, 1964.

[21] R.M. Bolle, N.K. Ratha, and S. Pankanti. Error analysis of pattern recognition systems-the subsets bootstrap. *Computer Vision and Image Understanding*, 93(1):1–33, 2004.

[22] J.G. Bosch, S.C. Mitchell, B.P.F. Lelieveldt, F. Nijland, O.Kamp, M. Sonka, and J.H.C. Reiber. Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Transactions on Medical Imaging*, 21(11):1374–1383, 2002.

[23] V. Bruce, A.M. Burton, and N. Dench. *What's distinctive about a distinctive face?* Lawrence Erlbaum Associates, 1988.

[24] V. Bruce, P.J.B. Hancock, and A.M. Burton. *Human face perception and identification*. Springer-Verlag, 1998.

[25] A.M. Burton, V. Bruce, and P.J.B. Hancock. Vital signs of identity. *Cognitive Science*, 23(1):1–31, 1999.

[26] C. Butakoff and A.F. Frangi. A framework for weighted fusion of multiple statistical models of shape and appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1847–1857, 2006.

[27] B.F. Buxton and M.B. Dias. Implicit, view invariant, linear flexible shape modelling. *Pattern Recognition Letters*, 26(4):433–447, 2005.

[28] P. Campadelli, R. Lanzarotti, G. Lipori, and E. Salvi. Face and facial feature localization. In *Proc. 13th Int. Conf. on Image Analysis and Processing, Cagliari, Italy. Lecture Notes in Computer Science vol. 3617*, pages 1002–1009, 2005.

[29] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: an approach based on registration of texture mapped 3D models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):322–336, 2000.

[30] A. Caunce and C.J. Taylor. Using local geometry to build 3D sulcal models. In *Proc. 16th Int. Conf. on Information Processing in Medical Imaging, Visegrád, Hungary. Lecture Notes in Computer Science vol. 1613*, pages 196–209, 1999.

[31] X. Chai, L. Qing, S. Shan, X. Chen, and W. Gao. Pose invariant face recognition under arbitrary illumination based on 3D face reconstruction. In *Proc. 5th Int. Conf. on Audio- and Video-Based Biometric Person Authentication, New York, NY, USA. Lecture Notes in Computer Science vol. 3546*, pages 956–965, 2005.

[32] V. Chatzis, A.G. Bors, and I. Pitas. Multimodal decision-level fusion for person authentication. *IEEE Transactions on Systems Man and Cybernetics Part A-Systems and Humans*, 29(6):674–680, 1999.

[33] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, 83(5):705–741, 1995.

[34] C. Chen, M. Zhao, S.Z. Li, and J. Bu. Parameter optimization for active shape models. In *Proc. 6th Asian Conf. on Computer Vision, Jeju Island, Korea*, volume 2, pages 1068–1073, 2004.

[35] O. Chum, T. Pajdla, and P. Sturm. The geometric error for homographies. *Computer Vision and Image Understanding*, 79(1):86–102, 2005.

[36] T.F. Cootes, D. Cooper, C.J. Taylor, and J. Graham. A trainable method of parametric shape description. *Image and Vision Computing*, 10(5):289–294, 1992.

[37] T.F. Cootes, G. Edwards, and C.J. Taylor. Active appearance models. In *Proc. 5th European Conf. on Computer Vision, Freiburg, Germany. Lecture Notes in Computer Science vol. 1407*, pages 484–498, 1998.

[38] T.F. Cootes, G. Edwards, and C.J. Taylor. A comparative evaluation of active appearance model algorithms. In *Proc. British Machine Vision Conf., Southampton, UK*, volume 2, pages 680–689, 1998.

[39] T.F. Cootes, G.J. Edwards, and C J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[40] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Comparing active shape models with active appearance models. In *Proc. British Machine Vision Conf., Nottingham, UK*, volume 1, pages 173–182, 1999.

[41] T.F Cootes, G.J. Page, C.B. Jackson, and C.J. Taylor. Statistical gray-level models for object location and identification. *Image and Vision Computing*, 14(8):533–540, 1996.

[42] T.F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision. Technical report, Wolfson Image Analysis Unit, University of Manchester, UK, 2001.

[43] T.F. Cootes and C.J. Taylor. Statistical models of appearance for medical image analysis and computer vision. In *Proc. SPIE Medical Imaging, San Diego, CA, USA*, volume 4332, pages 236–248, 2001.

[44] 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(1):38–59, 1995.

[45] T.F. Cootes, C.J. Taylor, and A. Lanitis. Multi-resolution search with active shape models. In *Proc. 12th Int. Conf. on Pattern Recognition, Jerusalem, Israel*, volume 1, pages 610–612, 1994.

[46] T.F. Cootes, G. Wheeler, K.N. Walker, and C.J. Taylor. Coupled-view active appearance models. In *Proc. British Machine Vision Conf., Bristol, UK*, volume 1, pages 52–61, 2000.

[47] T.F. Cootes, G. V. Wheeler, K.N. Walker, and C.J. Taylor. View-based active appearance models. *Image and Vision Computing*, 20(9):657–664, 2002.

[48] I.J. Cox, J. Ghosh, and P.N. Yianilos. Feature-based face recognition using mixture-distance. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Francisco, CA, USA*, pages 209–216, 1996.

[49] I. Craw, N. Costen, T. Kato, and S. Akamatsu. How should we represent faces for automatic recognition? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):725–736, 1999.

[50] D. Cristinacce and T.F. Cootes. A comparison of shape constrained facial feature detectors. In *Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Seoul, Korea*, pages 375–380, 2004.

[51] D. Cristinacce and T.F. Cootes. Feature detection and tracking with constrained local models. In *Proc. British Machine Vision Conf., Edinburgh, UK*, pages 929–938, 2006.

[52] C. Davatzikos, X. Tao, and D. Shen. Hierarchical active shape models, using the wavelet transform. *IEEE Transactions on Medical Imaging*, 22(3):414–423, 2003.

[53] R.H. Davies, C.J. Twining, T.F. Cootes, J.C. Waterton, and C.J. Taylor. 3D statistical shape models using direct optimisation of description length. In *Proc. 7th European Conf. on Computer Vision, Copenhagen, Denmark. Lecture Notes in Computer Science vol. 2352*, pages 3–20, 2002.

[54] M. de Bruijne, B. van Ginneken, W.J. Niessen, M. Loog, and M.A. Viergever. Model-based segmentation of abdominal aortic aneurysms in CTA images. In *Proc. SPIE Medical Imaging, San Diego, CA, USA*, volume 4684, pages 463–474, 2002.

[55] M. de Bruijne, B. van Ginneken, M.A. Viergever, and W. Niessen. Adapting active shape models for 3D segmentation of tubular structures in medical images. In *Proc. 18th Int. Conf. on Information Processing in Medical Imaging, Ambleside, UK. Lecture Notes in Computer Science vol. 2732*, pages 136–147, 2003.

[56] D. DeCarlo. *Generation, Estimation and Tracking of Faces*. PhD thesis, University of Pennsylvania, Department of Computer and Information Science, Philadelphia, PA, USA, 1998.

[57] G. Dedeoglu, S. Baker, and T. Kanade. Resolution-aware fitting of active appearance models to low resolution images. In *Proc. 9th European Conf. on Computer Vision, Graz, Austria. Lecture Notes in Computer Science vol. 3952*, pages 83–97, 2006.

[58] H. Demirel, T.J. Clarke, and P.Y.K. Cheung. Adaptive automatic facial feature segmentation. In *Proc. 2nd IEEE Int. Conf. on Automatic Face and Gesture Recognition, Killington, VT, USA*, pages 277–282, 1996.

[59] M.B. Dias and B.F. Buxton. Separating shape and pose variations. *Image and Vision Computing*, 22(10):851–861, 2004.

[60] I.L. Dryden and K. Mardia. *Statistical Shape Analysis*. Wiley, 1998.

[61] Y. Du and X. Lin. Multi-view face image synthesis using factorization model. In *Proc. ECCV 2004 Int. Workshop on Human-Computer Interaction, Prague, Czech Republic. Lecture Notes in Computer Science vol. 3058*, pages 200–210, 2004.

[62] N. Duta and M. Sonka. Segmentation and interpretation of MR brain images. an improved active shape model. *IEEE Transactions on Medical Imaging*, 17(6):1049–1067, 1998.

[63] G.J. Edwards, A. Lanitis, C.J. Taylor, and T.F. Cootes. Statistical models of face images - improving specificity. *Image and Vision Computing*, 16(3):203–211, 1998.

[64] E. Erzin, Y. Yemez, and A.M. Tekalp. Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia*, 7(5):840–852, 2005.

[65] L. Fan and K. Sung. Model-based varying pose face detection and facial feature registration in colour images. *Pattern Recognition Letters*, 24(1):237–249, 2003.

[66] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[67] L. Florack. *The Syntactical Structure of Scalar Images*. PhD thesis, Utrecht University, Utrecht, The Nedherlands, 2001.

[68] M. Frigge, D.C. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50–54, 1989.

[69] C.D. Frowd, P.J.B. Hancock, and D. Carson. EvoFIT: a holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Perception*, 1(1):19–39, 2004.

[70] A. Gee and R. Cipolla. Estimating gaze from a single view of a face. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Seattle, WA, USA*, pages 758–760, 1994.

[71] S.J. Gibson, C.J. Solomon, and A.P. Bejarano. Synthesis of photographic quality facial composites using evolutionary algorithms. In *Proc. British Machine Vision Conf., Norwich, UK*, pages 221–230, 2003.

[72] A.J. Goldstein, L.D. Harmon, and A.B. Lesk. Identification of human face. *Proceedings of the IEEE*, 59(5):748–760, 1971.

[73] D. González-Jiménez and J.L. Alba-Castro. Toward pose-invariant 2-D face recognition through point distribution models and facial symmetry. *IEEE Transactions on Information Forensics and Security*, 2(3):413–429, 2007.

[74] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society B*, 53(2):285–339, 1991.

[75] The International Biometric Group. Biometric market and industry report 2006-2010. www.biometricgroup.com, 2006.

[76] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7):1157–1182, 2003.

[77] J. Hager. *Asymmetries in Facial Expression; Emotion in the Human Face*. Cambridge Univ. Press, 1982.

[78] G. Hamarneh and T. Gustavsson. Deformable spatio-temporal shape models: Extending ASM to 2D+time. In *Proc. British Machine Vision Conf., Manchester, UK*, pages 13–22, 2001.

[79] G. Hamarneh and T. Gustavsson. Deformable spatio-temporal shape models: Extending active shape models to 2D+time. *Image and Vision Computing*, 22(6):461–470, 2004.

[80] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[81] C. Hoogendoorn, F.M. Sukno, S. Ordas, and A.F. Frangi. Bilinear models for spatio-temporal point distribution analysis: Application to extrapolation of whole heart cardiac dynamics. In *Proc. IEEE ICCV 2007 8th Int. Workshop on Mathematical Methods in Biomedical Image Analysis, Rio de Janeiro, Brazil*, pages 1–8, 2007.

[82] T. Horprasert, Y. Yacoob, and L.S. Davis. Computing 3D head orientation from a monocular image sequence. In *Proc. 2nd IEEE Int. Conf. on Automatic Face and Gesture Recognition, Killington, VT, USA*, pages 242–247, 1996.

[83] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for face recognition. In *Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Seoul, Korea*, pages 843–850, 2004.

[84] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[85] A.K. Jain, R. Bolle, and S. Pankani (eds). *Biometrics: Personal identificacion in a networked society*. Kluwer, New York, 1999.

[86] O. Jesorsky, K.J. Kirchberg, and R.W. Frischholz. Robust face detection using the hausdorff distance. In *Proc. 3th Int. Conf. on Audio- and Video-Based Biometric Person Authentication, Halmstad, Sweden. Lecture Notes in Computer Science vol. 2091*, pages 90–95, 2001.

[87] F. Jiao, S.Z. Li, H.Y. Shum, and D. Schurmans. Face alignment using statistical models and wavelet features. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Madison, WI, USA*, volume 1, pages 321–327, 2003.

[88] G.A. Jones and J.M. Jones. *Information and Coding Theory*. Springer, 2000.

[89] T. Kanade. *Picture Processing by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto University, Kyoto, Japan, 1973.

[90] A. Kanaujia and D. Metaxas. Large scale learning of active shape models. In *Proc. IEEE Int. Conf. on Image Processing, Austin, TX, USA*, volume 1, pages 265–268, 2007.

[91] H. Kang, T.F. Cootes, and C.J. Taylor. A comparison of face verification algorithms using appearance models. In *Proc. British Machine Vision Conf., Manchester, UK*, volume 2, pages 477–486, 2002.

[92] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[93] M.D. Kelly. *Model guided face recognition with multi-scale analysis*. PhD thesis, Stanford University, Stanford, CA, USA, 1970.

[94] S. Kim, S.T. Chung, S. Jung, D. Oh, J. Kim, and S. Cho. Multi-scale Gabor feature based eye localization. In *Proc. World Academy of Science, Engineering and Technology, Vienna, Austria*, volume 21, pages 483–487, 2007.

[95] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[96] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, and M.A. Abidi. Recent advances in visual and infrared face recognition, a review. *Computer Vision and Image Understanding*, 97(1):103–135, 2005.

[97] A. Koschan, S.K. Kang, J.K. Paik, B.R. Abidi, and M.A. Abidi. Video object tracking based on extended active shape models with color information. In *Proc. 1st European Conf. on Color in Graphics, Imaging, and Vision, Poitiers, France*, pages 126–134, 2002.

[98] M.A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *Archival Journal in Chemical Engineering*, 37(2):233–243, 1991.

[99] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Reliability-based decision fusion in multimodal biometric verification systems. *Journal on Advances in Signal Processing*, pages Article ID 86572, 9 pages, 2007.

[100] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1):25–41, 2000.

[101] B. Kurt, A.S. Etaner-Uyar, T. Akbal, N. Demir, A.E. Kanlikilicer, M.C. Kus, and F.H. Ulu. Active appearance model-based facial composite generation with interactive nature-inspired heuristics. In *Proc. Int. Workshop on Multimedia Content Representation, Classification and Security, Istanbul, Turkey. Lecture Notes in Computer Science vol. 4105*, pages 183–190, 2006.

[102] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[103] A. Lanitis, C.J. Taylor, T. Ahmed, and T.F. Cootes. Classifying variable objects using a flexible shape model. In *Proc. 5th Int. Conf. on Image Processing and its Applications, Edinburgh, UK*, pages 70–74, 1995.

[104] A. Lanitis, C.J. Taylor, and T.F. Cootes. Recognizing human faces using shape and gray level information. In *Proc. IEEE Int. Conf. on Automation, Robotics and Computer Vision, Singapore*, pages 1153–1157, 1994.

[105] A. Lanitis, C.J. Taylor, and T.F. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401, 1995.

[106] A. Lanitis, C.J. Taylor, and T.F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[107] R. Larsen, M.B. Stegmann, S. Darkner, S. Forchhammer, T.F. Cootes, and B.K. Ersboll. Texture enhanced appearance models. *Computer Vision and Image Understanding*, 106(1):20–30, 2007.

[108] K. Lekadir and G.Z. Yang. Carotid artery segmentation using an outlier immune 3D active shape models framework. In *Proc. 9th Int. Conf. Medical Image Computing and Computer Assisted Intervention, Copenhagen, Denmark. Lecture Notes in Computer Science vol. 4190*, pages 620–627, 2006.

[109] H. Li and O. Chutatape. Boundary detection of optic disk by a modified ASM method. *Pattern Recognition*, 36(9):2093–2104, 2003.

[110] K.P. Li and J.E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, New York, NY, USA*, pages 595–598, 1988.

[111] S.Z. Li and A.K. Jain. *Handbook of Face Recognition*. Springer, 2005.

[112] S.Z. Li, H. Zhang, Y.S. Cheng, and Q. Cheng. Multi-view face alignment using direct appearance models. In *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Washington, DC, USA*, pages 309–314, 2002.

[113] Y. Li, S. Gong, and H.M. Liddell. Modelling faces dynamically across views and over time. In *Proc. 8th IEEE Int. Conf. on Computer Vision, Vancouver, BC, Canada*, pages 554–559, 2001.

[114] Z. Li and H. Ai. Texture-constrained shape prediction for mouth contour extraction and its state estimation. In *Proc. 18th Int. Conf. on Pattern Recognition, Hong Kong, China*, volume 2, pages 88–91, 2006.

[115] L. Liang, F. Wen, X. Tang, and Y. Xu. An integrated model for accurate shape alignment. In *Proc. 9th European Conf. on Computer Vision, Graz, Austria. Lecture Notes in Computer Science vol. 3954*, pages 333–346, 2006.

[116] A.W.C. Liew, S.H. Leung, and W.H. Lau. Segmentation of color lip images by spatial fuzzy clustering. *IEEE Transactions on Fuzzy System*, 11(4):542–549, 2003.

[117] C. Liu, H.Y. Shum, and C.I Zhang. Hierarchical shape modeling for automatic face localization. In *Proc. 7th European Conf. on Computer Vision, Copenhagen, Denmark. Lecture Notes in Computer Science vol. 2351*, pages 687–703, 2002.

[118] C. Liu and H. Wechsler. A shape- and texture-based enhanced fisher classifier for face recognition. *IEEE Transactions on Image Processing*, 10(4):598–608, 2001.

[119] X. Liu, P. Tu, and F. Wheeler. Face model fitting on low resolution images. In *Proc. British Machine Vision Conf., Edinburgh, UK*, pages 1079–1088, 2006.

[120] E. Lleida and R.C. Rose. Utterance verification in continuous speech recognition: decoding and training procedures. *IEEE Transactions on Speech and Audio Processing*, 8(2):126–139, 2000.

[121] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[122] X. Lv and J. Zhou. Robust corner detection under varying illumination. In *Proc. Int. Conf. on Information Intelligence and Systems, Bethesda, MD, USA*, pages 396–398, 1999.

[123] N. Magnenat-Thalmann, H. Minh, M. de Angelis, and D. Thalmann. Design, transformation and animation of human faces. *The Visual Computer*, 51(1):32–39, 1989.

[124] B.S. Manjunath, R. Chellappa, and C. von der Malsburg. A feature based approach to face recognition. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Urbana-Champaign, IL, USA*, pages 373–378, 1992.

[125] A. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):748–763, 2002.

[126] A. Martínez and R. Benavente. The AR face database. Technical report, Computer Vision Center, Barcelona, Spain, 1998.

[127] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, N. Capdevielle, W. Gerstner, Y. Abdeljaoued, J. Bigun, S. Ben-Yacoub, and E. Mayoraz. Comparison of face verification results on the XM2VTS database. In *Proc. 15th Int. Conf. on Pattern Recognition, Barcelona, Spain*, volume 4, pages 858–863, 2000.

[128] I. Matthews, J. Xiao, and S. Baker. 2D vs. 3D deformable face models: Representational power, construction, and real-time fitting. *International Journal of Compuer Vision*, 75(1):93–113, 2007.

[129] T. Maurer and C. von der Malsburg. Single-view based recognition of faces rotated in depth. In *Proc. 2nd IEEE Int. Conf. on Automatic Face and Gesture Recognition, Killington, VT, USA*, pages 176–181, 1996.

[130] R. McGill, J.W. Tukey, and W.A. Larsen. Variations of boxplots. *The American Statistician*, 32(1):12–16, 1978.

[131] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marceland S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F.B. Tek, G.B. Akar, F. Deravi, and N. Mavity. Face verification competition on the XM2VTS database. In *Proc. 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication, Guildford, UK. Lecture Notes in Computer Science vol. 2688*, pages 964–974, June 9-11 2003.

[132] K. Messer, J.V. Kittler, J. Short, G. Heusch, F. Cardinaux, S. Marcel, Y. Rodriguez, S. Shan, Y. Su, W. Gao, and X. Chen. Performance characterisation of face recognition algorithms and their sensitivity to severe illumination changes. In *Proc. Int. Conf. on Biometrics, Hong Kong, China. Lecture Notes in Computer Science vol. 3832*, pages 1–11, 2006.

[133] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Proc. 2nd Int. Conf. on Audio- and Video-Based Biometric Person Authentication, Washington DC, USA*, pages 72–77, 1999.

[134] S.C Mitchell, B.P. Lelieveldt, R.J. van der Geest, H.G. Bosch, J.H. Reiber, and M Sonka. Time-continuous segmentation of cardiac MR image sequences using active appearance motion models. In *Proc. SPIE Medical Imaging, San Diego, CA, USA*, volume 4322, pages 249–256, 2001.

[135] S.C. Mitchell, B.P.F. Lelieveldt, R.J. van der Geest, H.G. Bosch, J.H.C. Reiver, and M. Sonka. Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac MR images. *IEEE Transactions on Medical Imaging*, 20(5):415–423, 2001.

[136] T. Mitchell and M. Sarhadi. Non-linear statistical models for the 3D reconstruction of human body pose and motion from monocular image sequences. *Image and Vision Computing*, 18(9):729–737, 2000.

[137] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[138] Y. Moses, Y. Adini, and S. Ullman. Face recognition: the problem of compensating for illumination changes. In *Proc. 3rd European Conf. on Computer Vision, Stockholm, Sweden. Lecture Notes in Computer Science vol. 800*, pages 286–296, 1994.

[139] M.M. Nordstrom, M. Larsen, J. Sierakowski, and M.B. Stegmann. The IMM face database: An annotated dataset of 240 face images. Technical report, Denmark, 2004.

[140] E.J. Ong and S. Gong. Tracking hybrid 2D-3D human models from multiple views. In *Proc. IEEE Int. Workshop on Modelling People, Kerkyra, Greece*, pages 11–18, 1999.

[141] S. Ordas, E. Oubel, R. Leta, F. Carrera, and A.F. Frangi. A statistical shape model of the heart and its application to model-based segmentation. In *Proc. SPIE Medical Imaging, San Diego, CA, USA*, volume 6511, 2007.

[142] S. Ordas, H.C. van Assen, L. Boisrobert, M. Laucelli, J. Puente, B.P.F. Lelieveldt, and A.F. Frangi. Statistical modeling and segmentation in cardiac MRI using a grid computing approach. In *Proc. European Workshop on Advances on Grid Computing, Amsterdam, The Netherlands. Lecture Notes in Computer Science vol. 3470*, pages 6–15, 2005.

[143] A. Ortega, F.M. Sukno, E. Lleida, A.F. Frangi, A. Miguel, L. Buera, and E. Zacur. AV@CAR: A spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In *Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisbon, Portugal*, volume 3, pages 763–767. (www.cilab.upf.edu/ac), 2004.

[144] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 22(12):1424–1445, 2000.

[145] V. Perlibakas. Distance measures for PCA-based face recognition. *Pattern Recognition Letters*, 25(6):711–724, 2004.

[146] D. Perperidis. *Spatio-temporal registration and modelling of the heart using cardiovascular MR imaging*. PhD thesis, Department of Computing, Imperial College London, UK, 2005.

[147] P.J. Phillips, D. Blackburn, P. Grother, E. Newton, and J.M. Bone. Methods for assessing progress in face recognition. Biometric Systems: Technology, Design and Performance Evaluation, London, UK, 2003.

[148] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. Evaluation Report, http://www.frvt.org, 2003.

[149] P.J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[150] P.J. Phillips, P. Rauss, and S. Der. FERET (face recognition technology) recognition algorithm development and test results. Technical report, Army Research Laboratory, USA, 1996.

[151] 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.

[152] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Detection of interest points using symmetry. In *Proc. 3rd IEEE Int. Conf. on Computer Vision, Osaka, Japan*, pages 62–65, 1990.

[153] D. Riccio and J.C. Dugelay. Geometric invariants for 2D/3D face recognition. *Pattern Recognition Letters*, 28(14):1907–1914, 2007.

[154] M. Rogers and J. Graham. Exploiting weak shape constraints to segment capillary images in microangiopathy. In *Proc. 3rd Int. Conf. Medical Image Computing and Computer Assisted Intervention, Pittsburgh, PA, USA. Lecture Notes in Computer Science vol. 1935*, pages 717–726, 2000.

[155] M. Rogers and J. Graham. Robust active shape model search. In *Proc. 7th European Conf. on Computer Vision, Copenhagen, Denmark. Lecture Notes in Computer Science vol. 2353*, pages 289–312, 2002.

[156] G.L. Rogova and V. Nimier. Reliability in information fusion: Literature survey. In *Proc. 7th Int. Conf. on Information Fusion, Stockholm, Sweden*, volume 2, pages 1158–1165, 2004.

[157] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error function. In *Proc. 7th European Conf. on Computer Vision, Copenhagen, Denmark. Lecture Notes in Computer Science vol. 2353*, pages 3–19, 2002.

[158] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel PCA. In *Proc. British Machine Vision Conf., Nottingham, UK*, pages 483–492, 1999.

[159] S. Romdhani, A. Psarrou, and S. Gong. On utilising template and feature-based correspondence in multi-view appearance models. In *Proc. 6th European Conf. on Computer Vision, Dublin, Ireland. Lecture Notes in Computer Science vol. 1842*, pages 799–813, 2000.

[160] A. Ross and A.K. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, 2003.

[161] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Willey, New York, 1987.

[162] D. Ruprecht and H. Muller. Image warping with scattered data interpolation. *IEEE Computer Graphics and Applications*, 15(2):37–43, 1995.

[163] H. Sackheim, R.C. Gur, and M.C. Saucy. Emotions are expressed more intensively on the left side of the face. *Science*, 202(4366):434–436, 1978.

[164] A. Samal and P.A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1):65–77, 1992.

[165] F.S. Samaria and A.C. Harter. Parameterization of a stochastic model for human face identification. In *Proc. 2nd IEEE Workshop on Applications of Computer Vision, Sarasota, FL, USA*, pages 138–142, 1994.

[166] C. Sanderson, S. Bengio, and Y. Gao. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288–302, 2006.

[167] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[168] S. Sclaroff and J. Isidoro. Active blobs. In *Proc. 6th IEEE Int. Conf. on Computer Vision, Bombay, India*, pages 1143–1153, 1998.

[169] I.M. Scott, T.F. Cootes, and C.J. Taylor. Improving appearance model matching using local image structure. In *Proc. 18th Int. Conf. on Information Processing in Medical Imaging, Ambleside, UK. Lecture Notes in Computer Science vol. 2732*, pages 258–269, 2003.

[170] M. Seise, S.J. McKenna, I.W. Ricketts, and C.A. Wigderowitz. Learning active shape models for bifurcating contours. *IEEE Transactions on Medical Imaging*, 26(5):666–677, 2007.

[171] A. Selinger and D. Socolinsky. Appearance-based facial recognition using visible and thermal imagery: a comparative study. Technical report, Equinox Corporation, http://www.equinoxsensors.com, 2002.

[172] P.P. Smyth, C.J. Taylor, and J.E. Adams. Vertebral shape automatic measurement with active shape models. *Radiology*, 211(2):571–578, 1999.

[173] J.M. Sotoca, J.M. Inesta, and M.A. Belmonte. Hand bone segmentation in radioabsorptiometry images for computerised bone mass assessment. *Computerized Medical Imaging and Graphics*, 27(6):459–467, 2003.

[174] P. Sozou, T.F. Cootes, C.J. Taylor, and E. DiMauro. Non-linear point distribution modelling using a multi-layer perceptron. In *Proc. British Machine Vision Conf., Birmingham, UK*, pages 107–116, 1995.

[175] S. Srisuk, M. Petrou, W. Kurutach, and A. Kadyrov. Face authentication using the trace transform. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Madison, WI, USA*, pages 305–312, 2003.

[176] F.M. Sukno and A.F. Frangi. Exploring reliability for automatic identity verification with statistical shape models. In *Proc. IEEE Workshop on Automatic Identification Advanced Technologies, Alguero, Italy*, pages 80–86, 2007.

[177] F.M. Sukno, J.J. Guerrero, and A.F. Frangi. Homographic active shape models for view-independent facial analysis. In *Proc. SPIE Biometric Technologies for Human Identification, Orlando, FL, USA*, volume 5779, pages 152–163, 2005.

[178] F.M. Sukno, S. Ordas, C. Butakoff, S. Cruz, and A.F. Frangi. Active shape models with invariant optimal features (IOF-ASMs). In *Proc. 5th Int. Conf. on Audio- and Video-Based Biometric Person Authentication, New York, NY, USA. Lecture Notes in Computer Science vol. 3546*, pages 365–375, 2005.

[179] F.M. Sukno, S. Ordas, C. Butakoff, S. Cruz, and A.F. Frangi. Active shape models with invariant optimal features: Application to facial analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1105–1117, 2007.

[180] J. Sung, T. Kanade, and D. Kim. A unified gradient-based approach for combining ASM into AAM. *International Journal of Compuer Vision*, 75(2):297–309, 2007.

[181] B. Takacs and H. Wechsler. Detection of faces and facial landmarks using iconic filter banks. *Pattern Recognition*, 30(10):1623–1636, 1997.

[182] T. Tamminen and J. Lampinen. A bayesian occlusion model for sequential object matching. In *Proc. British Machine Vision Conf., Kingston, UK*, volume 2, pages 547–556, 2004.

[183] T. Tamminen and J. Lampinen. Sequential Monte Carlo for bayesian matching of objects with occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):930–941, 2006.

[184] H.H. Thodberg and A. Rosholm. Application of the active shape model in a commercial medical device for bone densitometry. In *Proc. British Machine Vision Conf., Manchester, UK*, volume 1, pages 43–52, 2001.

[185] C.E. Thomaz and D.F. Gillies. A new fisher-based method applied to face recognition. In *Proc. 10th Int. Conf. on Computer Analysis of Images and Patterns, Groningen, The Netherlands. Lecture Notes in Computer Science vol. 2756*, pages 596–605, 2003.

[186] Y. Tong, Y. Wang, Z. Zhu, and Q. Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 40(11):3195–3208, 2007.

[187] Z. Tu and S.C. Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):657–673, 2002.

[188] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[189] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.

[190] H.C. van Assen, M.G. Danilouchkine, A.F. Frangi, S. Ordás, J.J.M. Westenberg, J.H.C. Reiber, and B.P.F. Lelieveldt. SPASM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data. *Medical Image Analysis*, 10(2):286–303, 2006.

[191] B. van Ginneken, A.F. Frangi, J.J. Staal, B.M. ter Haar Romeny, and M.A. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, 2002.

[192] B. van Ginneken, M.B. Stegmann, and M. Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006.

[193] P.F. Velleman and D.C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press, 1981.

[194] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[195] K.N. Walker, T.F. Cootes, and C. J. Taylor. Correspondence using distinct points based on image invariants. In *Proc. British Machine Vision Conf., Essex, UK*, volume 1, pages 540–549, 1997.

[196] K. Wan, K. Lam, and K. Ng. An accurate active shape model for facial feature extraction. *Pattern Recognition Letters*, 26(15):2409–2423, 2005.

[197] J.G. Wang, E. Sung, and R. Venkateswarlu. EM enhancement of 3D head pose estimated by perspective invariance. In *Proc. ECCV 2004 Int. Workshop on Human-Computer Interaction, Prague, Czech Republic. Lecture Notes in Computer Science vol. 3058*, pages 187–199, 2004.

[198] W. Wang, S. Stan, W. Gao, B. Cao, and B. Yin. An improved active shape model for face alignment. In *Proc. 4th IEEE Int. Conf. on Multimodal Interfaces, Pittsburgh, PN, USA*, pages 523–528, 2002.

[199] T. Wilhelm, H.J. Böhme, and H.M. Gross. Classification of face images for gender, age, facial expression, and identity. In *Proc. 15th Int. Conf. on Artificial Neural Networks: Biological Inspirations, Warsaw, Poland. Lecture Notes in Computer Science vol. 3696*, pages 569–574, 2005.

[200] L. Wiskott, J.M Fellows, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[201] C.B.H. Wolstenholme and C.J. Taylor. Wavelet compression of active appearance models. In *Proc. 2nd Int. Conf. Medical Image Computing and Computer Assisted Intervention, Cambridge, UK. Lecture Notes in Computer Science vol. 1679*, pages 544–554, 1999.

[202] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D active appearance models. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Washington, DC, USA*, volume 2, pages 535–542, 2004.

[203] S. Xin and H. Ai. Face alignment under various poses and expressions. In *Proc. 1st Int. Conf. on Affective Computing and Intelligent Interaction, Beijing, China. Lecture Notes in Computer Science vol. 3784*, pages 40–47, 2005.

[204] S. Yan, X. He, Y. Hu, Y. Zhang, M. Li, and Q. Cheng. Bayesian shape localization for face recognition using global and local textures. *IEEE Transactions on Circuits and Systems For Video Technology*, 14(1):102–113, 2004.

[205] S. Yan, M. Li, H. Zhang, and Q. Cheng. Ranking prior likelihood distributions for bayesian shape localization framework. In *Proc. 9th IEEE Int. Conf. on Computer Vision, Nice, France*, 2003.

[206] S. Yan, C. Liu, S. Li, H. Zhang, H. Shum, and Q. Cheng. Texture-constrained active shape models. In *Proc. ECCV 2002 1st Int. Workshop on Generative-Model Based Vision, Copenhagen, Denmark*, volume 2, pages 1031– 1034, 2002.

[207] S. Yan, C. Liu, S.Z. Li, H. Zhang, H.Y. Shum, and Q. Cheng. Face alignment using texture-constrained active shape models. *Image and Vision Computing*, 21(1):69–75, 2003.

[208] S.C. Yan, X. Hou, S.Z. Li, H.J. Zhang, and Q.S. Cheng. Face alignment using view-based direct appearance models. *International Journal of Imaging Systems and Technology*, 13(1):106–112, 2003.

[209] A.L. Yuille, P. Halliman, and D.S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.

[210] L. Zhang. Estimation of eye and mouth corner point positions in a knowledge-based coding system. In *Proc. SPIE Digital Compression Technologies and Systems for Video Communications, Berlin, Germany*, volume 2952, pages 21–28, 1996.

[211] L. Zhang, H. Ai, and S. Lao. Robust face alignment based on hierarchical classifier network. In *Proc. ECCV 2006 Int. Workshop on Human-Computer Interaction, Graz, Austria. Lecture Notes in Computer Science vol. 3979*, pages 1–11, 2006.

[212] L. Zhang, H. Ai, S. Xin, C. Huang, S. Tsukiji, and S. Lao. Robust face alignment based on local texture classifiers. In *Proc. IEEE Int. Conf. on Image Processing, Genoa, Italy*, volume 2, pages 354–357, 2005.

[213] Z. Zhangl, L. Zhu, S.Z. Li, and H. Zhang. Real-time multi-view face detection. In *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Washington, DC, USA*, pages 142–147, 2002.

[214] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[215] Y. Zhou, L. Gu, and H. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Madison, WI, USA*, volume 1, pages 109–116, 2003.

[216] Y. Zhou, W. Zhang, X. Tang, and H. Shum. A bayesian mixture model for multi-view face alignment. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA*, volume 2, pages 741–746, 2005.

Publications

Peer reviewed papers in international journals

- **F.M. Sukno**, S. Ordas, C. Butakoff, S. Cruz, and A.F. Frangi. Active shape models with invariant optimal features: Application to facial analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1105-1117, 2007.
- **F.M. Sukno** and A.F. Frangi. Reliability Estimation for Statistical Shape Models. *Conditionally accepted for publication in IEEE Transactions on Image Processing, pending minor revision*
- **F.M. Sukno**, J.J. Guerrero and A.F. Frangi. Projective Active Shape Models for Pose-variant Image Analysis of Quasi-Planar Objects: Application to Facial Analysis. *Submitted for publication*
- C. Hoogendoorn, **F.M. Sukno**, S. Ordas, and A.F. Frangi. Bilinear Models for Spatiotemporal Point Distribution Analysis: Application to Extrapolation of Left Ventricular, Biventricular and Whole Heart Cardiac Dynamics. *Submitted for publication*

Peer reviewed papers in conference proceedings

- C. Hoogendoorn, **F.M. Sukno**, S. Ordas, and A.F. Frangi. Bilinear models for spatiotemporal point distribution analysis: Application to extrapolation of whole heart cardiac dynamics. In *Proc. IEEE ICCV 2007 8th Int. Workshop on Mathematical Methods in Biomedical Image Analysis, Rio de Janeiro, Brazil*, pages 1-8, 2007.
- **F.M. Sukno** and A.F. Frangi. Exploring reliability for automatic identity verification with statistical shape models. In *Proc. IEEE Workshop on Automatic Identification Advanced Technologies, Alguero, Italy*, pages 80-86, 2007.
- D. González-Jiménez, **F.M. Sukno**, J.L. Alba-Castro and A.F. Frangi. Automatic pose correction for local feature-based face authentication. In *Proc. 4th IEEE Workshop on*

Motion of Non-Rigid and Articulated Objects, Mallorca, Spain. Lecture Notes in Computer Science vol. 4069, pages 356-365, 2006.

- **F.M. Sukno**, S. Ordas, C. Butakoff, S. Cruz, and A.F. Frangi. Active shape models with invariant optimal features (IOF-ASMs). In *Proc. 5th Int. Conf. on Audio- and Video-Based Biometric Person Authentication, New York, NY, USA. Lecture Notes in Computer Science vol. 3546*, pages 365-375, 2005.
- **F.M. Sukno**, J.J. Guerrero and A.F. Frangi. Homographic active shape models for view-independent facial analysis. In *Proc. SPIE Biometric Technologies for Human Identification, Orlando, FL, USA*, volume 5779, pages 152-163, 2005.
- A. Ortega, **F.M. Sukno**, E. Lleida, A.F. Frangi, A. Miguel, L. Buera, and E. Zacur. AV@CAR: A spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In *Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisbon, Portugal*, volume 3, pages 763-767. (www.cilab.upf.edu/ac), 2004.
- A. Ortega, **F.M. Sukno**, E. Lleida, A.F. Frangi, A. Miguel, L. Buera, and E. Zacur. Base de datos audiovisual y multicanal en castellano para reconocimiento automático del habla multimodal en el automv́il. In *III Jornadas en Tecnologías del Habla*, pages 125-130, (www.cilab.upf.edu/ac), 2004.

Resumen

La presente tesis se centra en la aplicación de métodos estadísticos al modelado y análisis facial. A partir del paradigma de los modelos de forma y apariencia, introducidos en la última década, se proponen extensiones algorítmicas que permiten mejorar su confianbilidad y su invarianza a distintos tipos de rotaciones. Dichas extensiones se plantean de forma genérica, tal que los modelos conserven su relevancia original en otros ámbitos de aplicación.

La validación experimental de las técnicas propuestas se ha realizado en torno al análisis facial. Este campo ha cobrado especial relevancia en los últimos años, con un importante crecimiento de su mercado a nivel internacional. Uno de los hechos más destacables es la reciente adopción de biometría facial como tecnología de base en los nuevos pasaportes, relegando así a un segundo plano a otras modalidades como la basada en las huellas dactilares y en el iris, a pesar de ofrecer estas últimas menores tasas de error. Siendo la forma natural de identificación en el ser humano, el reconocimiento facial se ha visto favorecido por ser percibido como menos intrusivo que otras modalidades y por ofrecer, en teoría, la posibilidad de operar sin necesidad de colaboración de la persona a identificar.

En el Capítulo 1 se hace una breve introducción a la biometría y se describen los componentes básicos de un sistema automático de reconocimiento facial, contextualizando así los modelos de forma y apariencia. Concretamente, los modelos activos de forma (ASM por sus siglas en inglés) constituyen el principal componente metodológico de esta tesis y su teoría se introduce someramente en el Capítulo 2.

Los ASM permiten la segmentación y análisis automático de imágenes a través de un método generativo. Introducidos en 1992 (por T. Cootes et al.), han dado lugar a numerosas publicaciones centradas en la aplicación de los ASM a diversos tipos de imágenes, entre las cuales las imágenes médicas y faciales han sido las más comunes, pero no las únicas. Por otro lado, el modelado por medio de ASM resultó ser excesivamente simplificado para algunas aplicaciones. Como resultado, una segunda línea de publicaciones se concentra actualmente en extensiones y mejoras a la formulación original. Una de las más importantes dió origen en 1998 a los modelos activos de apariencia (AAM). De hecho, los AAM pronto tomaron entidad propia, de forma independiente a los ASM.

La presente tesis introduce tres nuevas extensiones a los ASM que se centran en mejorar dichos models en cuanto a su invarianza y confiabilidad. La hipótesis de trabajo es que la mejora de los algoritmos de segmentación derivará en una delineación más precisa de los rasgos faciales, permitiendo una interpretación más adecuada de la información contenida en las imágenes faciales.

La primera de estas extensiones enfoca el problema de segmentar con precisión los distintos rasgos faciales en tomas frontales (Capítulo 3). El método propuesto generaliza los ASMs utilizando un modelo de intensidad no lineal para la descripción de la imagen a nivel local. Dichos descriptores son invariantes a transformaciones rígidas y su número puede controlarse mediante algoritmos de selección secuencial. Al contrario que en la formulación original de los ASM, la distribución de los valores de intensidad en el conjunto de entrenamiento no se asume unimodal ni Gaussiana. El método propuesto ha demostrado una mejora significativa en la precisión de las segmentaciones con respecto a los ASM, lo que también se ha visto reflejado en menores tasas de error en experimentos de verificación de identidad.

La segunda extensión (Capítulo 4) se centra en la invarianza del algoritmo de segmentación en presencia de rotaciones fuera del plano de la imagen cuando se trabaja con objetos casi-planos. Acotando adecuadamente las regiones faciales analizadas, es posible asumir que sus contornos son aproximadamente coplanares y utilizar entonces conceptos de geometría proyectiva. De este modo, los ASM se modifican de modo tal que su funcionamiento sea más robusto ante rotaciones fuera del plano (siempre que los contornos permanezcan visibles). Como resultado, un ASM construido sólo con tomas frontales puede utilizarse para la segmentación de imágenes tomadas desde otros ángulos. El mátodo propuesto se ha validado en imágenes faciales seleccionadas sistemáticamente de acuerdo al ángulo de rotación fuera del plano, incluyendo tres vistas hacia cada lado más dos vistas con variaciones arriba-abajo. Estos experimentos constituyen la mayor evaluación cuantitativa en segmentación bajo variaciones sistemáticas de pose facial publicada hasta la fecha.

La tercera extensión (Capítulo 5) provee una medida de confianza para el análisis efectuado sobre cada imagen. Es decir, el modelo pueda estimar automáticamente si el resultado de una segmentación es confiable o no. Esto se vuelve muy importante cuando los ASM son utilizados en sistemas totalmente automáticos, donde el éxito de la etapa de segmentación es crucial para la posterior interpretación de la imagen. Esta estimación de confiabilidad puede ser de relevancia en varias aplicaciones. Como ejemplo se han abordado dos de ellas: selección automática de modelos y verificación confiable de identidad, obteniendo en ambos casos resultados altamente positivos. La principal fortaleza del método propuesto reside en su bajo índice de falsos positivos (las imágenes incorrectamente segmentadas rara vez son clasificadas como confiables).

De este modo, las dos primeras extensiones comparten el concepto de invarianza (a rotaciones en y fuera del plano de la imagen). Por otro lado, se verá que la primera extensión también incrementa la precisión de las segmentaciones, mientras que la tercera extensión se concentra en la estimación del grado de éxito obtenido en el procesado de cada una de las imágenes. En los tres casos se ha realizado una exhaustiva valiación experimental, demostrando la superioridad de las técnicas propuestas sobre otras técnicas existentes en la literatura.

Acknowledgements

I would like to thank several people who have contributed to the accomplishment of this thesis. In first place to my advisor, Dr. Alejandro Frangi, for his continuous commitment to this work, both from the scientific and organizational aspects which have been fundamental. I am also very grateful to all other co-authors of the different parts of this thesis, Constantine Butakoff, Santiago Cruz, Josechu Guerrero and Sebastián Ordás, with whom I have had very pleasant collaboration and useful discussions.

I would also like to thank all my lab colleagues throughout these five years, for the friendly work environment they contributed to create, and very specially to the closer ones working in Biometrics and in the Statistical Image Analysis group. And as well to my former colleagues of GTC in Zaragoza, especially to those who collaborated in the hard work for the acquisition of the AV@CAR database, part of which has served as test material for this work. Indeed, almost half of my PhD took place at Centro Politécnico Superior in University of Zaragoza, certainly one of the most friendly environments I have been working in.

Thanks also to Prof. Eduardo Lleida Solano, who helped me a lot in the first period of my PhD and accepted later to be my tutor to facilitate my continuity in the PhD program at University of Zaragoza regardless of the transfer of the lab to Barcelona.

I am especially grateful to Dr. Carlos H. Muravchik, for guiding me in the selection of my current research group, to Milton Hoz de Vila for all kinds of help with logistics, software, etc. (including the cover of this thesis) and to the administrative staff of University of Zaragoza (at the Department of Electrical Engineering and Communications, the Aragon Institute of Engineering Research, Relaciones Internacionales and Tercer Ciclo) and Universitat Pompeu Fabra (at the Department of Information and Communication Technologies) for their kind support.

I would like to express my gratitude to the financial support provided by a number of entities: The Banco Santander grants program at University of Zaragoza, which not only allowed me to start my PhD studies but also generously contributed to the printing expenses of this thesis together with Scati Labs S.A. from Zaragoza, a company that has closely collaborated with us and encouraged our research in Biometrics; and to the Spanish Ministry

of Education and Science, Spanish Ministry of Industry and the European Commission, that have contributed by financing a number of research projects related to this thesis.

Finally, a special thanks to my parents, Alicia and Mateo, my brothers, Josefina, Serenella and Ricardo, and to my girlfriend, Chong, for their constant and unconditional encouragement and support.

Curriculum Vitae

Federico Mateo Sukno was born in 1975 in Rosario, Argentina. He received secondary education in National Technical School (E.N.E.T) number 2 from Neuquén, Argentina, obtaining the technician diploma in electronics in 1993; and in Romania, within the Argentine's Air Force distance learning program Condor (Bachiller, 1993).

He graduated in Electrical Engineering at the Universidad Nacional de La Plata, Argentina in 2000. After graduation he worked for two years in mobile telecommunications (Ericsson) and industrial control (ABB); both in Argentina.

In 2003 he moved to Spain and enrolled in the Biomedical Engineering PhD program at the University of Zaragoza, with a grant from this University in collaboration with Banco Santander.

He is currently working on statistical shape models in the Research Group for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Information and Communication Technologies, Pompeu Fabra University (in Barcelona), where he is also an Assistant Professor in coding and information theory and analog electronics.