Aspects of facial biometrics for verification of personal identity

M. Ulises Ramos Sánchez

Submitted for the Degree of Doctor of Philosophy from the University of Surrey

Centre for Vision, Speech and Signal Processing
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

May 2000

© M. Ulises Ramos Sánchez 2000
Summary

This thesis studies various aspects of facial biometrics for the verification of personal identity in a multimodal framework. The research focuses on the mouth area and, more specifically, on the design of a lip tracking system for the extraction of visual features.

The tracker is based on statistical chromaticity models and uses a B-spline representation of the contour of the lips. Shape variability is restricted to affine deformations of a linear combination of modes of shape variation, which are automatically estimated in a robust way using the tracking results provided by a first, rather unconstrained lip tracker.

Tracking experiments were performed in a large multimedia database and the results were fed as input features to a Dynamic Time Warping algorithm for speaker verification purposes according to a published evaluation protocol. A weighted linear classifier is eventually trained for performing fusion experiments on the same database combining various verification modalities such as face and voice.

Key words: Biometrics, Personal Identity Verification, Face Recognition, Lip Reading, Lip Tracking, Fusion.

Email: m.ramos-sanchez@ee.surrey.ac.uk
WWW: http://www.eim.surrey.ac.uk/
Acknowledgements

I would like to thank my supervisor, Josef Kittler, for his guidance and suggestions during the course of this research, and very specially for his favourable attitude and support taking into account the added difficulties of supervising work that has been carried out mostly on a part-time basis. I am very much indebted to Yongping Li for his invaluable collaboration and helpfulness, specially during the last, quite involving lip tracking experiments. Kieron Messer’s help with video format conversion is greatly acknowledged. Many thanks go to Medha Pandit for providing the DTW software that was adapted for the verification experiments, to Martin Dobsik for his collaboration in updating the lip tracking software, and to José Luis Fernández for his useful advice regarding the applicability of various optimisation algorithms. Jiri Matas’s guidance during the first year and his many useful suggestions are greatly appreciated. Most of the software developed builds upon the existing, home-made GAMA library; the efforts of current and former Group members involved in its development are therefore gratefully acknowledged. I am also grateful to Jürgen Lüttin. His research on lip tracking has been an excellent reference for this work, and his openness to discuss aspects of it is greatly appreciated. Very special thanks to Mike Whybray, my former Group Leader in BT Labs, for his support during the nearly two years I spent in Martlesham. The time spent at the Labs has been instrumental for this research work, not only for the direct experience of working on face recognition or the financial support during those two years, which is gratefully acknowledged, but also for the possibility of benefiting from stimulating discussion with people like Gareth Edwards and Chris Taylor, University of Manchester; Baback Moghadam and Sandy Pentland, MIT Media Lab; John Daugman, University of Cambridge; Andrew Bangham, Stephen Cox and Iain Matthews, University of East Anglia; and Bill Welsh, Mark Shackleton, Ian Scott, Charley Nightingale, Ed Munday, Pat Mulroy and Nigel Garnham, BT Labs. The funding of the M2VTS Project during the first year of this research is also acknowledged. Many thanks as well to the M2VTS partners for making their verification results available, which have been used in this thesis in the course of the final fusion experiments reported.
Contents

1 Introduction
   1.1 Information fusion ........................................ 2
   1.2 Research objectives ...................................... 3
   1.3 Thesis overview .......................................... 4

2 Face recognition .............................................. 7
   2.1 Introduction .............................................. 7
   2.2 Face recognition by humans .............................. 8
   2.3 Face localisation ......................................... 10
      2.3.1 Colour .............................................. 10
      2.3.2 Texture ............................................ 11
      2.3.3 Motion: image differencing ......................... 11
      2.3.4 Context: facial feature grouping ................. 12
      2.3.5 Model-based approaches ......................... 12
   2.4 Facial feature extraction ............................... 13
      2.4.1 Intensity distribution ................................ 13
      2.4.2 Physical properties ................................ 13
      2.4.3 Colour ............................................. 14
      2.4.4 Depth .............................................. 14
      2.4.5 Correlation-based approaches .................... 14
      2.4.6 Deformable templates .............................. 14
      2.4.7 Dynamic contours: snakes ......................... 15
   2.5 Descriptors and matching techniques .................. 15
      2.5.1 Geometrical features vs templates .......... 15
2.5.2 Robust correlation ........................................ 16
2.5.3 Spatial frequencies ........................................ 16
2.5.4 Eigenfaces ................................................. 17
2.5.5 Active Shape Models ....................................... 18
2.5.6 Local Feature Analysis ..................................... 19
2.5.7 Facial deformations ........................................ 19
2.5.8 Linear discriminant techniques ............................ 20
2.5.9 Probabilistic matching ...................................... 21
2.5.10 Active Appearance Models ............................... 22
2.5.11 Hidden Markov Models .................................... 22
2.5.12 Neural networks .......................................... 22
2.6 Addressing face variability as a 3D object ............... 23
2.6.1 Varying pose ............................................... 23
2.6.2 Profiles vs frontal views .................................. 24
2.7 Discussion .................................................... 25

3 Lip-reading ................................................... 27
3.1 Introduction .................................................. 27
3.2 Lip-reading by humans ....................................... 28
3.3 Automatic lip-reading for speech recognition ............ 29
3.4 Lip modelling and tracking .................................. 30
3.4.1 Deformable templates ..................................... 31
3.4.2 Mouth region descriptors .................................. 31
3.4.3 Contour-based approaches ............................... 32
3.4.4 Other approaches ......................................... 32
3.5 Dynamic contours for lip tracking ......................... 33
3.5.1 Deterministic physically-based dynamic models ....... 34
3.5.2 B-spline tracking with Kalman filters ................. 35
3.5.3 Refinements and extensions .............................. 37
3.6 Discussion .................................................... 38
Contents

4 The bootstrap tracker 41

4.1 Introduction 41
4.2 Statistical colour modelling 42
  4.2.1 Extraction of colour models 43
4.3 Lip representation 45
4.4 Spline initialisation 46
4.5 Lip boundary estimation 47
  4.5.1 Estimation based on log-likelihood ratio 47
  4.5.2 Enhancing robustness of the lip boundary estimation 49
4.6 Tracking mechanism 52
  4.6.1 Estimation with spatial and temporal smoothing 52
  4.6.2 An alternative approach based on weighted least squares 54
4.7 Tracking results 57
4.8 Discussion 58

5 Robust estimation of main modes of lip shape variation 63

5.1 Introduction 63
5.2 Robust covariance estimation 63
5.3 Minimum Covariance Determinant Estimator 65
5.4 MCD for robust eigenlip estimation 66
5.5 Experimental results 68
5.6 Eigennips computation 69
5.7 Summary 72

6 Shape-constrained lip tracking 75

6.1 Introduction 75
6.2 Shape models 75
6.3 Dynamics model and tracking mechanism 77
6.4 Tracking results 79
6.5 More complex dynamics models? 79
6.6 Discussion 84
6.7 Summary 85
## Contents

7 Application of lip-based features to the verification of personal identity 87

7.1 Introduction .................................. 87
7.2 Description of the XM2VTS database and the Lausanne protocol 88
7.3 Extraction of lip features and matching strategy ............... 88
7.4 Lip-based verification of personal identity .................... 91
7.5 Speaker verification results on the XM2VTS database according to the Lausanne protocol .......... 94
7.6 Fusion experiments on the XM2VTS database according to the Lausanne protocol ....... 94
7.7 Conclusions ................................... 97

8 Conclusions ................................ 99
8.1 Research contributions ......................... 99
8.2 Future work .................................. 101

A Robust clustering .................................. 103
A.1 Introduction .................................. 103
A.2 Generalised minimum covariance determinant clustering ....... 103
A.3 Fuzzy version .................................. 105
A.4 Experimental results and discussion ..................... 106

B Automatic lip tracking initialisation ......................... 109
B.1 Introduction .................................. 109
B.2 Face location and approximation by an elliptical shape ........ 110
B.3 Generation of lip candidates based on grey-level gradient projection ...... 111
B.4 Selecting the best candidate ......................... 112
B.5 Discussion .................................... 113

C Robust facial characterisation and reconstruction .......... 115
C.1 Introduction .................................. 115
C.2 Face space modelling ............................ 116
C.3 Reconstruction from incomplete measurements .............. 117
    C.3.1 Sample-based least squares estimation ............... 117
Chapter 1

Introduction

Personal identity verification is an issue for a number of applications. It is required for controlled entry to secure sites, it is an essential step in performing financial transactions, and it starts to be of common use for secure access to teleservices, teleshopping and telebanking. The undeniable fact that biometric signatures are something intrinsic to the individual (their applicability being only limited by their different discriminatory information and the state of the art of technology to make the best of them), without requiring the user to memorise any "secret" code or having to hold any external security token (e.g. smart cards, keys, etc), the power of which lies on preventing their access to any person other than the intended user, makes them no doubt very attractive as authentication means.

For this reason a number of automatic personal identity verification methods have been or are being developed. These include the traditional approaches such as online signature verification, finger print recognition, and voice recognition (see e.g. [1]) as well as more recent modalities such as hand geometry identification, iris identification and face recognition.

Some of the modalities available would not be acceptable for general user access applications. (The simplicity of the verification procedure and its unobtrusive nature would eliminate many candidates which in their own right may be considerably more reliable). This means that other alternatives may have to be considered which are less powerful and cannot achieve reliable verification of personal identity on their own.

There are two main directions along which some of the shortcomings of existing user friendly verification approaches can be addressed. In the first instance the discriminative power of monomodal techniques can be enhanced by means of differential verification analysis which is aimed at disambiguating user personal identity signatures (in whatever modality) using the most discriminative attributes.

The second approach is to explore the possibility of multimodal verification, involving the personal identity signatures derived from more than one verification modality. This second approach was the subject of the M2VTS Project of the ACTS Research and Technology Development Programme [2].
The main objective of the M2VTS Project was to extend the scope of application of network-based services by adding novel and intelligent functionalities, enabled by automatic verification systems combining multimodal strategies (secured access based on speech, image and other information). The objectives were also to show that the limitations of individual technologies can be overcome by relying on multimodal decisions (combination or fusion of these technologies) and can find practical and important applications in the new emerging fields of advanced interfaces for tele-services.

1.1 Information fusion

The need for information fusion has been identified in many research areas. It has been shown [3] that multiple cues play a crucial role in image interpretation and that the use of interpretation strategies which depend on the image data, temporal context and visual goals significantly simplifies the complexity of the image interpretation problem and makes it computationally feasible. The approach was applied to a vision system that combined shape, colour, motion, prior scene knowledge and object motion behaviour is described.

More related to our research area, person identification systems based on acoustic and visual features had already been described by Brunelli et al [4, 5]. In [4] the system is organized as a set of non-homogeneous classifiers whose outputs are integrated after a normalization step. In particular, two classifiers based on acoustic features and three based on visual ones provide data for an integration module whose performance is evaluated. HyperBF networks are used for the integration of multiple classifiers at an hybrid rank/measurement level. The performance of the integrated system is shown to be superior to that of the acoustic and visual subsystems. Vector quantisation is used to design static and dynamic codebooks of a given reference speaker. The utterances used are sequences of seven Italian digits in a continuous way and in whatever order. Face recognition performs template matching [6] on frontal, expressionless images with minor modifications. The overall identification time is 5 seconds.

Another system based on geometrical facial features and achieving similar results can be found in [5]. Although geometrical features performance for face recognition was reported slightly worse than the obtained using template matching [6], face representation is more compact and less computationally demanding, thus supplying an example of integration where simple and fast recognition modules are combined.

Within the frame of the M2VTS Project some early attempts at information fusion were reported in [7, 8]. In [7], a multimodal person verification system using voice and images is proposed. The visual part currently involves i) matching of a coarse grid containing Gabor phase information from face images, ii) facial feature localisation and extraction iii) 3D biometrical feature extraction by structured light. The acoustic part uses three methods (Dynamic Time Warping, SOSM -for Second Order Statistical Method- and Hidden Markov Models) to compare voice references extracted from the speech signal. In the acoustic part linear prediction coefficients
1.2. Research objectives

are extracted and three different classifiers are used in parallel. The global decision is taken by applying a Furui threshold to the individual methods and in combining the individual results according to a majority law. In [8] fusion experiments are reported with reference to the three voice recognition methods above on a 40-person database.

More recently, and as a spin-off of the M2VTS project, a large multimedia database was created, the so-called XM2VTS database [9], and a protocol for evaluation of multimodal verification systems and information fusion algorithms in that database was published -the so-called Lausanne Protocol [10]. Recent verification results on the XM2VTS database according to the Lausanne Protocol are described in [11], where a number of voice and facial verification modalities are combined in a multimodal fashion to generate a multimodal system that outperforms its facial and vocal modules when taken separately.

1.2 Research objectives

Originated in the framework of the M2VTS project, the generic research objective of this thesis was the development of enabling techniques for the interpretation of facial imagery orientated to the multimodal verification of person identity.

More specifically, the research focused on the mouth area, with the aim of extracting image information that could lead to the development of a speaker verification system based on visual features, and to its integration with other verification modalities.

An obvious subsidiary task to accomplish beforehand, and which eventually attracted most of the attention, was the development of suitable tracking techniques for extracting visual features from the mouth area. A survey of the prior art in lip tracking showed that certain approaches relied on the use of lip highlighters [12], whilst others based their success on limiting the degrees of freedom of some sort of parametric representation of the lips contour (e.g. [13, 14, 15]). In a personal identity verification scenario where it is of the utmost importance to preserve as much identity-related variability as possible, some of the approaches were found to limit shape variability too strongly from the very beginning, or the procedure followed to obtain more general shape modes of variation (for instance, by manually annotating landmarks in a sufficiently large number of training images) was found neither adequate nor of practical use in more realistic scenarios. Therefore, the development of a lip tracker that relied on a minimum number of working assumptions and constraints, and that were able to automatically learn any such shape models became a research issue.

Taking into account reported and experienced difficulties with grey-level images, the use of colour information and in particular the robust extraction of colour models was to be investigated in order to build the tracking system within a sound statistical framework.
1.3 Thesis overview

Chapter 2 is a survey of current facial analysis, localisation and recognition techniques. Some of the topics presented include the perception and recognition of faces by humans, face localisation, extraction of facial features, descriptors and matching techniques and face variability as a 3D object. Although these aspects may appear somewhat peripheral to the main stream of the thesis, some of them impact directly on the development of this lip-based speaker verification since, for instance, the lips still have to be located in the image, which is usually easier once the face has been. Some problems and techniques are therefore quite related and thus this chapter also serves the purpose of depicting the background against which this research has been carried out.

Chapter 3 is concerned with the interpretation of visual information from the mouth area, commonly referred to as lip-reading. The use of visual stimuli for speech interpretation seems to be a general characteristic of human perception, not exclusive of hearing-impaired people. The use of lip-reading for speech recognition is overviewed and so will be lip-modelling and tracking, with special emphasis on dynamic contours.

Three chapters have been devoted to the developed lip tracking system. Chapter 4 is concerned with the so-called bootstrap tracker, a colour-based tracking system with hardly any shape constraint but the smoothness imposed by the B-spline model representation.

In chapter 5, it is shown how the tracking results of the bootstrap tracker can be used to robustly estimate in an automatic way the most statistically representative modes of shape variation. The method represents a novel contribution of robust statistics to the computer vision field. A robust covariance matrix estimator, Rousseeuw's MCD algorithm [16] is tailored in order to obtain a robust, low-dimensional characterisation of the space of lip shapes, thus overcoming the effect of incorrect tracking results in their computation.

Chapter 6 presents the final version of the lip tracking system, built upon the estimation of main modes of shape variation described in the preceding chapter 5. Since the main modes of shape variation are given by the eigenvectors of the robust covariance matrix estimation, the tracking system is termed an eigenlip tracker where shape variation is limited to affine deformations of linear combinations of those shape variation modes -the so-called eigenlips. A simple first order dynamics model is adopted whereby the estimate for the current frame is used as a prediction for the following one. Tracking proceeds by finding the optimal eigenlip components and the parameters of an affine transform that best match the extracted measurements.

The application of the tracking results to lip-based personal identity verification within the framework of the Lausanne Protocol are presented in chapter 7. Visual features are used in a text dependent verification system based on a Dynamic Time Warping algorithm. Results are shown in the XM2VTS database for both configurations of the Lausanne protocol. Finally, fusion experiments using a weighted linear classifier incorporating information from diverse combinations of verification modalities are described as well.
Accurate colour modelling is of central importance for satisfactory tracking performance. Appendix A describes a clustering algorithm that was used for extracting the statistical chromaticity models used by the final version of the lip tracker. A 'crisp' version and a fuzzy one are presented, the latter performing better in cases of poor separability between clusters. The optimal number of clusters is also automatically determined as part of the algorithm.

Building upon the robust estimation developments described in chapter 5 and Appendix A a procedure was devised to help automate lip localisation as much as possible with a view to the realisation of the quite involving tracking experiments on the XM2VTS database. The method, that combines statistical colour and shape models, as well as grey level information, is described in Appendix B.

Robust parameter estimation has been a recurrent theme throughout this research work whether for colour modelling, extraction of modes of shape variation, tracking or lip initialisation. A seminal work that proved instrumental for establishing the parallelism between these problems and eventually coming to statistically sound solutions was the robust regression problem encountered in a different, though related, context, namely the estimation of the coefficients of the expansion in eigenfaces of facial images affected by occlusion. This piece of work is described in Appendix C.
Chapter 2

Face recognition

2.1 Introduction

Automatic face recognition has many applications, ranging from static matching of controlled photographs to analysis of surveillance video images. The applications have different constraints in terms of complexity, processing requirements and, accordingly, present a wide range of different technical challenges. The subject has attracted researchers in many different areas, such as engineering, biometrics, psychology, etc., but, as claimed in [17], little synergism seems to exist between the studies in psychophysics and the engineering literature.

Psychological studies of human face recognition suggest that virtually every type of available information is used [6]. Another interesting point is that face recognition is possible even at coarse resolution [18], which implies that the overall geometrical configuration of the face features would be sufficient for discrimination. However, the techniques applied are not only global, and, as proved by the results reported in e.g. [6] or [19] among others, feature analysis contributes to enhance the performance of those global systems, what means that, however sufficient, more information is still available and could be exploited.

Apart from the intrinsic problem of establishing which features should be extracted to perform face recognition at its best, several factors contribute to make face recognition more difficult. Faces are 3D objects but we will usually have to deal with 2D images of them. The images, however, convey additional variability to that we could expect associated with each different individual. Although the list does not intend to be exhaustive nor rigorous, typical factors influencing face variability are scale, pose, lighting conditions, expression, aging, occlusion, noise, etc. Scale will appear as variations in image size, e.g. associated with variations in the distance to the camera. Pose stands for viewing geometry, which combined with illumination can cause undesirable shading effects. Illumination variability represents intensity as well as number and direction of the lighting sources. The subject’s expression also plays its role in face variability, and to some considerable extent with a certain personal character. Aging, as well as other shorter term appearance changes over time, should be borne in mind. Occlusion means that we cannot guarantee that
access to the proper face will always be available. In some cases, the presence of obscuring objects can be expected. Finally, noise would encompass other factors that have not been specifically considered, such as the sensor characteristics, etc.

This chapter has been organised in the following way. First, face recognition by humans is reviewed as a starting point. Next, current techniques for face localisation and facial feature extraction are addressed in sections 2.3 and 2.4 respectively, followed by an analysis of typical descriptors and matching techniques as applied for face recognition in section 2.5. The variability of faces in connection with their 3D nature is addressed in section 2.6 which is followed by a final discussion section.

2.2 Face recognition by humans

Tovee [20] refers to human faces as a social semaphore and stresses the remarkable transformation undergone by primate faces. The extensive increase in their innervation, musculature and flexibility from the almost rigid, mask-like faces of some New World monkeys to the flexible, highly mobile face of the great apes reaches its maximum sophistication in humans. It is pointed out that this increasing sophistication does not simply respond to identification purposes. As primates developed into more complex social groups the primate faces would develop into a kind of semaphore system, capable of signalling a wide range of complex social information, the recognition and interpretation of these social cues becoming extremely important for the smooth functioning of a social group and for an individual's place within the group hierarchy. Along these lines, Daugman [21] refers to the general consensus in behavioural biology that the main factor driving the evolution of large brains in primates was the computational load of this 'social' processing. The increasing complexity of facial innervation and musculature would have been paralleled by an increasing sophistication of the neural representation of facially signalled information [22].

The central role of face recognition in human life, and the seemingly effortless way humans do it, have motivated an interest in studying the mechanisms of human facial recognition, as can be seen, for instance, in Baron's theory for human visual information [23], which analysed the mapping from retina to cortex, and hypothesised a neural network topology for carrying out face recognition and explaining some typical syndromes.

Several researchers (e.g. [24]) have referred to the human ability of recognising faces as being 'hard-wired' in the brain. Bichsel [18] reports how, only 9 minutes after its birth, a baby is already capable of tracking face-like objects, which would point to an innate ability of the human brain. However, most of the overwhelming human face recognition abilities seem to be learned, according to the following summary of developmental studies [18].

After 2 weeks a baby is able to recognise its mother's face and looks at human faces in preference to other stimuli. Under 10 years, children seem to use only piecemeal information to identify faces; after this age, they develop the ability to
2.2. Face recognition by humans

use configurational information, which would account for the improved capacity of older children to encode unfamiliar faces in terms of the properties that reflect identity across various changes in the stimulus (e.g. clothing, hairstyle, lighting and viewpoint among others). Face recognition accuracy with upright faces improves sharply between ages 6 and 10, whilst the performance on inverted faces remains constant, which would suggest an adaptation of the recognition strategy to the typical format in which they see the faces in their society.

About the claimed difficulties in discriminating the faces of other races, there is evidence that the facial features appropriate for discriminating among black faces are different from those appropriate for white faces. For black faces, texture and color of the skin, face shape, thickness of the lips and breadth of the nose seem to be important features, whereas for white faces length, texture and colour of the hair, eye shape, fatness of the face and skin texture are more salient features. Attempts to train adult people to obtain better cross-race recognition accuracy are reportedly of little success, therefore concluding that the adaptation to faces of our own race also occurs during infancy.

The existence of face cells in the brains of the primates is apparently well established [22]. These are cells specially responsive to faces that show virtually no response to any other stimulus tested. Face cells are sensitive to the relative position of features within the face (the interocular distance is particularly important, as is the distance from eyes to mouth, and the amount and style of hair on the forehead). Face cells also continue to respond to VDU (video display unit) images of faces that have been low- or high-pass filtered so that they have no spatial frequencies in common, that have had the colour removed or altered or that have had the contrast reduced to a very low level. Similarly, the presentation of a single facial component elicits only a fraction of the response generated by the whole face, and removal of a single component of a face reduces, but does not eliminate, the response of a cell to a face.

Most face cells are selective for the viewing angle, but another small proportion would be responsive irrespective of the viewing angle. Similar findings about the orientation selectivity of temporal lobe face-selective cells in monkey brains are also reported in [25].

Face-selective neurons appear to be members of ensembles for coding faces rather than individual face detectors or grandmother cells. These so-called grandmother cells were being described as being the top of a processing pyramid that began with line and edge detectors in the striate cortex and continued with detectors of increasing complexity until a unit was reached that represented one specific object or person. Although individual cells respond differently to different faces, there is no evidence for a face cell that responds exclusively to one individual face [22]. It seems, rather, that face cells would comprise a distributed network where visual stimuli are encoded by the pattern of activity across the ensemble of neurons [25, 26]. According to [22], the responses of face-selective cells appear relatively strongly tuned and, as a result, face encoding would be rather sparse, suggesting that the cell populations or ensembles may be as small as a hundred neurons.

The left and the right hemispheres process face information differently [22] and it
seems that right hemisphere damage may be sufficient to cause prosopagnosia\(^1\). Presentation of faces to the left visual field (right hemisphere) leads to faster recognition than presentation to the right visual field (left hemisphere). The right-hemisphere advantage disappears when faces are presented upside down, and right-side damage disrupts recognition of upright faces but not inverted faces. It seems that, in the right hemisphere, upright faces are processed in terms of their feature configuration, whereas inverted faces are processed in a piecemeal manner, feature by feature -the brain would be mentally rotating the image [18]. In the left hemisphere, both upright and inverted faces seem to be processed in a piecemeal manner.

There seems to be strong evidence [22] for a disassociation between recognition of facially conveyed information, such as identity and emotion, and the interpretation of this information. The amygdala seems to be important both to interpret and to give meaning to the basic emotions in facial expressions, and seems to facilitate the differentiation of the blends of multiple emotions that the human face can signal.

With reference to the most important face region for recognition, psychological studies [18] confirm the popularly extended belief that this is the region containing the eyes. In fact, for many of the so called face cells, the presence of the eyes was essential for a response, whereas a lower number of cells responded selectively to the mouth region.

### 2.3 Face localisation

Face localisation is concerned with determining the position of the face in a more or less approximate way. Very often it is associated with face segmentation, where the face would be pulled out of the rest of the image. It is typically the first step in many face recognition tasks, because subsequent processing can be greatly simplified if restricted to a region of interest that is assumed to be the face. However, there are approaches that ignore this prior step and approaches that take place before face localisation, e.g. when salient features inside the face are localised first.

The availability of suitable data grabbing facilities can greatly favour subsequent processing for face recognition tasks and has been considered by some authors. This is the approach followed in [27], where a ‘silicon retina’ is presented. The silicon retina camera is a simplified VLSI model of the neural circuitry in the retina. Its utilisation ensures enhanced robustness against illumination changes, thus easing subsequent face localisation.

### 2.3.1 Colour

Face colour classification (FCC) is described in [28] for face localisation and tracking. FCC classifies input image pixels into two classes: skin and non-skin, using

---

\(^1\)Prosopagnosia is a selective impairment restricted to the recognition of faces, although no thoroughly studied prosopagnostic has yet been found to be free from associated deficits. As face-selective cells do not form a majority in any known site, lesions which include them might not be expected to produce a pure face agnosia in monkeys or humans [25].
2.3. Face localisation

the chromatic distribution of the skin colour. Generic models of skin chromaticity are obtained by smoothing the chromaticity of 5 faces. Examples of FCC show a rather low number of false negatives (the skin is correctly classified in the face), but significant number of false positives (parts of both background and hair are classified as skin). Motion (simple differencing) and shape are used to eliminate false regions.

A similar method for face and mouth localisation is accomplished by colour histogram matching [29]. The dependence of colour histogram on viewpoint is probably not a problem as only frontal views are used. In [30], the orange-like parts of the image are enhanced by utilising the I component of the YIQ colour system. The images are smoothed and then a threshold is imposed to the I colour component.

Region segmentation based on hue-saturation characteristics has also been reported [31, 32, 33] for face localisation. Following colour segmentation, a connected component analysis is done to extract face candidates. The connected components are fitted with an ellipse computed according to the moments method. The quality of fit is measured and the best candidate is selected as face hypothesis that will be tracked over time.

2.3.2 Texture

Texture has been considered for extracting facial parts from low resolution images (faces of 20 × 20 pixels) in [30]. Three measures of textural features are used taking into account the definition of space gray-level dependence (SGLD) matrices: inertia, inverse difference and correlation. An array of measurements is obtained for each measure consisting of values obtained for different pixel separations. A set of inequalities is defined in terms of inertia, inverse difference and correlation. Wherever they hold it is assumed to be a face.

2.3.3 Motion: image differencing

Detection and tracking of the head location is important for model-based coding purposes. In [34], data difference images are obtained by subtracting consecutive low-pass spatio-temporal subbands and thresholded to produce binary images.

In [35] a hierarchical scheme to extract facial features and model the face is presented. The silhouette is found by histogram thresholding of absolute frame differences plus spatiotemporal filtering. Then the head is found by exploiting neck concavities through computation of the complex hull of the silhouette.

In [36] as well, image differencing plus thresholding is proposed to localise the face. Then the motion of the blobs in the resulting binary motion image is analysed. Head position is determined after application of a few simple rules. The size of the blob that is assumed to be the moving head determines the size of the subimage to send to the recognition stage.
2.3.4 Context: facial feature grouping

Context proves very useful to establish the goodness of facial feature findings. In [37] a system is described that uses facial feature grouping to localise faces. Coarse features that remain invariant under different viewpoints and orientations are searched for by means of Gaussian derivative filters such as bar (line) detectors using a steerable-scalable decomposition. When the correct size and orientation are found, many high responses are obtained. Under weak-perspective assumption, it is possible to obtain three affine invariants, so that some constraints can be imposed on partial face groups (PFGs) used to represent the face. Each detected feature is replaced with a line corresponding to the longitudinal axis of the feature. Constraints are verified for each pair of line segments and then for each PFG. A belief network (directed acyclic graph) is constructed for face modelling. PFGs are used as evidences to determine the probability of the presence of a face. System performance relies very much on accurate feature localisation. Promising results are reported with simple background, but the performance drops on complex background because of poor contrast, shadows, background clutter and scarcity of features needed to group a PFG (e.g. thin or no eyebrows at all).

2.3.5 Model-based approaches

Despite the successful results reported by the techniques reviewed so far, based on the exploitation of simple cues like colour, motion or texture, the reported success depends very much on the controllability of the working environment and, in order to enhance the robustness of face localisation it is common to resort to a combination of those single strategies. The cues used rapidly supply indicators of possible face locations but, without a more valid model of face-likeness to validate the hypotheses, false positives are more than a mere possibility.

Other approaches, more statistically sound, aim at modelling faces as instantiations of a class of objects. This poses the problem of finding characteristics or features which are generic for all faces (generality), whilst at the same time allowing for the distinction of a given face (specificity). Typical statistical characterisation of faces are based on Principal Component Analysis (PCA) and, accordingly, a good number of eigenrepresentations in different flavours are described in the literature (e.g. [38, 36, 39, 40]).

When using eigenrepresentations (see section 2.5), the distance from the feature space (DFFS) [41] can be interpreted as an estimate of a marginal component of the probabilistic density of the object in image space. A complete estimate must incorporate a second marginal density based on a complementary "distance in-feature-space" (DIFS). Arbitrarily complex densities can be modelled using a mixture of Gaussians, whose parameters are solved by using the Expectation Maximisation algorithm. The density estimate will then be used to compute a local measure of target saliency at each spatial position and at each given scale (multiscale saliency maps). This approach has been tested on human face images, applied to face and feature localisation, and on cluttered scenes showing different hand gestures.
2.4 Facial feature extraction

Building upon previous work [39], Active Appearance Models (AAM) [42] are presented as a straightforward approach for photo-realistic model matching [43]. The AAM contains a statistical model of the shape and grey-level appearance of faces which can generalise to most valid examples. Localising a face in an image involves finding model parameters which minimise the difference between the image and a synthesised model example, projected into the image.

2.4 Facial feature extraction

For some applications the focus is on facial features rather than on the face as a whole. Quite often, face localisation represents a prior stage in order to ease feature location and extraction, but there also exist other approaches where these features are directly sought after.

2.4.1 Intensity distribution

The eyes are found in [44] exploiting the fact that the iris appears dark and surrounded by a light sclera. The flow field around an iris radiates outwards. Lines are drawn at each edge point along the direction opposite the gradient. These will pass through several bins among a 2-dimensional array of accumulators. Image quality and illumination conditions can severely affect this approach. Also, specularities may intrude. For example, in some cases, glass specularities are detected as the proper eye location.

The particular intensity minimum occurring between the lips is exploited in [45] for the detection and tracking of the mouth. The mouth is described by this valley contour which is based between the lips. The contour is shown to exist independently of illumination, viewpoint, identity, and expression.

In a different approach [31, 32], facial feature extraction takes advantage of the darkness of eye sockets and mouth region. Two methods are proposed for feature localisation following dark region enhancement by means of morphological operations (i.e. erosion with a rectangular window). A watershed algorithm is compared with a min-max analysis. The watershed algorithm does not need preprocessing, although the definition of the threshold it uses is critical. It also shows a local scope, failing in some cases to completely merge the lips in one of the test sequences. The min-max method localises the position of the eyes and mouth by searching for intensity minima in grey-level profiles. This method has a global scope but requires orientation normalisation.

2.4.2 Physical properties

Symmetry properties of the face can be used to support face recognition schemes. For instance, a robust facial feature detector based on a generalized symmetry interest operator is presented in [46]. No special tuning is required if the face occupies 15-60
% of the image. The operator was tested on a large face data base with a success rate of over 95%. It assumes the presence of a single face in the image and claims to detect the centers of the eyes and the mouth for a wide range of sizes, rotations and lighting conditions.

In [35], the eyes and the mouth are localised taking into account eye-to-eye distance and vertical symmetry characteristics as well as geometrical constraints derived from statistical biometrics inferred from a training set.

2.4.3 Colour

Facial feature search can take advantage of special colour characteristics. Thus, further to localising the face, the approach described in [29] is also able to localise the mouth following colour histogram matching [47]. Another method for finding facial features is presented in [48]. The method uses simple single-band manipulations in the HLS colour space; in case of failure a more advanced method is used.

2.4.4 Depth

Depth information is used in [49] to improve the accuracy and reliability of feature localisations. First the facial structure (depth) is recovered by sampling pairs of face images from a video sequence (a camera is mounted on a tripod with wheels and moved by hand in front of a seated person). The background is automatically excluded by the match validation process because of the ambiguities arisen when attempting to match a mostly uniform image region. A disparity map is obtained and the face is extracted by thresholding such a map. Following facial depth recovery, the main inner facial features can be localised. The nose is found through correlation with a nose depth template on the processed disparity image to obtain a nose map. The nose location would be given by the centroid of the nose blob. The rest of the features are found by means of standard grey-level manipulation techniques.

2.4.5 Correlation-based approaches

Gray-level correlation is one of the most straightforward approaches for facial feature localisation and is reported, for instance, in [29] for eye localisation. Template matching, based on the same principle, i.e. correlation with a reference image called the template, has been used by other researchers, e.g. [6]. Simple though, plain correlation is sensitive to factors like intensity changes, scale and is computationally expensive.

2.4.6 Deformable templates

Deformable templates [50] are currently used for object description and recognition. They are reported in [51] for the localisation of iris centers, so that the face rotation angle can be found and the templates readjusted. In [50], deformable templates
2.5. Descriptors and matching techniques

are used to localise lips and eyes. Latching onto the appropriate object involves minimising an energy function that is a combination of terms due to valley, edge, peak, image and internal potentials. The coefficients of the cost function terms are varied according to their relative importance during the matching process.

2.4.7 Dynamic contours: snakes

Snakes (energy minimising contours, [52]) are used in [53] for facial feature extraction algorithms. Whilst the contours of the eye and mouth can be suitably modelled by deformable templates because of their analytically describable shapes, the shapes of the eyebrows, nostril and face are difficult to model using a deformable template and snakes work better. Experiments were done on frontal views of 12 models against a black background. The feature contours were extracted for each portrait. It seems to overcome tilting and no user involvement is required, although lighting of the faces must be controlled.

2.5 Descriptors and matching techniques

2.5.1 Geometrical features vs templates

Much of the early work in automatic face recognition focused on detecting individual features such as the eyes, nose, mouth, and head outline, and defining a face model by the position, size and relationships among these features [54].

In [6] a first technique based on the computation of a set of geometrical features [55] (nose width and length, mouth position, and chin shape), and a second one based on grey-level template matching are compared. Tests are run on a common database (frontal images of faces of 47 people). For both methods the images are normalised for scale, rotation and position by setting the intraocular distance. Image intensity is normalised by dividing each pixel by the average intensity in the neighbourhood. A Gaussian pyramid of prenormalised images is built and correlation computed hierarchically.

Matching based on geometrical features assumes the features have Gaussian distributions with different means but identical covariance matrix for different persons. As far as template matching is concerned, correlation scores for three templates (both eyes, nose, mouth) are computed and added. Template matching is quite different from straight grey-level correlation. It uses preprocessing of the image, transforming it into a map of the magnitude of the gradient, quite close therefore to an edge map.

The results obtained on the testing sets (about 90% correct recognition using geometrical features and perfect recognition using template matching) favour the implementation of the template-matching approach, although the authors themselves consider these findings specific to the task and to their implementation.
2.5.2 Robust correlation

Provided lighting variance can be compensated for, correlation represents a straightforward approach for facial matching, specially for solving the verification problem, where a probe face is compared against a target template corresponding to the claimed identity.

In [56] a method is proposed for fast face localisation and verification (identification) based on a robust form of correlation. Face matching is performed by maximising correlation in a search space determined by image affine transforms (to deal with geometrical face appearance variation due to changes in scale, rotation and translation) and photometric linear mappings (linear mapping between the grey levels in the reference and test images). A robust kernel using a quadratic function is implemented to examine correlation over the set of random points drawn using Sobol sequences. This Monte-Carlo technique speeds the evaluation of correlation approximately twenty five times and makes the optimisation process near-real time. In recognition experiments performed on the M2VTS database [57], the optimised robust correlation outperformed two standard techniques based on the Dynamic Link Architecture [58].

2.5.3 Spatial frequencies

In [59] matching pursuit filters are presented for face recognition. They consist of an adapted wavelet decomposition used to encode local information at multiple scales. The eye and nose regions are used because of being the most stable and least varying facial features. The filter for a particular feature is designed from examples from images in the gallery. The locations are marked by a human operator. The detection of the face is accomplished by running the interior face filter on a decimated version of the probe image. The search for facial features is guided by a priori knowledge of the geometry of the face and the estimated location of the center of the face. A further example of the application of wavelet decomposition, in this case for teleconferencing can be found in [60].

A Gabor-based method known as Dynamic Link Architecture (DLA) and implementing elastic matching is described in [58]. In a particular implementation, the image consists of a 2D array of nodes (vertices) that are labelled with a collection of features that describe the grey-level distribution. Edges joining the vertices are labelled with metric information on the relative position of vertices. Vertex labels are based on Gabor wavelet decomposition. Only the magnitude is considered to avoid oscillatory edge response. A local description of an image is created by sampling the wavelet decomposition at 5 logarithmically spaced frequency levels and 8 orientations. Their magnitudes form a feature vector. Edge labels are the Euclidean distance vectors between the vertices.

Elastic matching amounts to searching for a set of vertex positions which optimises a cost function consisting of a linear combination of an edge term (penalised by a rigidity parameter) and a vertex term. The process is done in two stages:
2.5. Descriptors and matching techniques

1. The image graph is shifted while keeping its form rigid till a minimum is reported.

2. Simulated annealing follows to permit small graph distortions and diffusion of the vertices.

Further examples of DLA implementations can be found in [61, 27, 62, 63].

2.5.4 Eigenfaces

Sirovich and Kirby pioneered the use of Principal Component Analysis for the characterisation of human faces [38], attaining a low dimensional representation system based on the eigenvectors ('eigenpictures') of the covariance matrix of an ensemble of face pictures.

Turk and Pentland describe in [36] a near-real-time system that can locate and track a subject's head, and then recognise the person by comparing characteristics of the face to those of known individuals. The approach treats the face recognition problem as an intrinsically two-dimensional (2D) recognition problem rather than requiring recovery of three-dimensional geometry, taking advantage of the fact that faces are normally upright and thus may be described by a small set of 2D characteristic views. The system functions by projecting face images onto a feature space that spans the significant variations among known face images. The significant features are known as 'eigenfaces', because they are the eigenvectors (principal components) of the set of the faces. The projection operation characterises an individual face by a weighted sum of the eigenface features, and so to recognise a particular face it is necessary only to compare these weights to those of known individuals. Some particular advantages of the approach are that it provides for the ability to learn and later recognise new faces in an unsupervised manner, and that it is easy to implement using a neural network architecture.

In addition to [36], experiments with eigenfaces for recognition and interactive search in a large-scale face database are presented in [19]. Accurate visual recognition is demonstrated using a database of thousands of faces. The problem of recognition under general viewing orientation is also examined. A view-based multiple-observer eigenspace technique is proposed for use in face recognition under variable pose. In addition, a modular eigenspace description technique is used which incorporates salient features such as the eyes, nose and mouth, in an eigenfeature layer. This modular representation yielded higher recognition rates as well as a more robust framework for face recognition. Eigenfeatures seem to overcome shortcomings of the standard eigenface method when facing situations such as hats, beards, etc. Image registration can be done automatically by detecting the most face-like object in the image and normalising for translation, scale and slight rotations, so as to align the position of both eyes [41]; following masking of the background, the image is eventually normalised for contrast.

The eigenface approach has been followed by other researchers as well. In [64] shape-free faces are matched by means of PCA analysis. Twenty out of fifty eigenfaces
obtained from a pool of images not present in the test pool are used. This method was compared with the dot product of shape-free representation. The results were significantly better using the eigenface representation, which the authors explain in terms of better generalisation capability by using a few codes chosen for their ability to describe variability between faces.

A testbed used to investigate different codings for automatic face recognition is described in [65]. Following manual location of 34 landmarks, a number of tests were conducted considering the intensity pattern of shape-free (i.e. landmarks aligned to a reference face) and correctly shaped faces, and the configuration of the landmarks themselves.

The PCA coding of shape-free faces was more effective than the corresponding coding of correctly shaped faces. And matching based on Mahalanobis distance proved more effective than using Euclidean distance. According to the authors, this clear advantage provides evidence that PCA is a more appropriate method of coding faces than simply using raw images; and that something more sophisticated than template matching is occurring. Manipulation within the coding to emphasise distinctive features of the faces, by caricaturing, allowed further increases in performance. Configuration also proved an effective method of recognition, with rankings given to incorrect matches relatively uncorrelated with those from shape-free faces. It is suggested that an appropriate combination of shape and texture parameters may be more effective than either system. Taken together, the results strongly support the suggestion that faces should be considered as lying in a high-dimensional manifold which is linearly approximated by these two factors, possibly with a separate systems for local features.

2.5.5 Active Shape Models

Active Shape Models (ASM) (see e.g. [67, 68]) are statistical models of local intensity distribution. ASMs rely on representing objects by sets of labelled points called landmarks. The landmarks are described by Point Distribution Models (PDM) derived by examining the statistics of the intensity values of their associated profiles. ASMs are mainly intended for object localisation and modelling. In addition to this, the PDMs representing each landmark augment the model and have also been successfully used for face recognition following localisation. Therefore, ASMs contain three layers of information: a) shape (determined by the position of the landmark points), b) intensity distribution associated with each PDM, and c) shape-free (by deforming each face to the mean shape) grey-level information.

Fitting a shape instance to an image is done following an iterative procedure by calculation of the point displacement and adjustments to the pose and shape parameters, i.e. find translation, rotation and scaling which best maps the instance displacement. Residual adjustments are only achieved by deforming the shape of the model. In [69] a further refinement consisting of a multiresolution approach is presented. A set of grey-level models for each landmark is constructed, one for

---

2 More details can be found in [66].
2.5. Descriptors and matching techniques

every level on a multi-resolution image pyramid. The same number of pixels is used at each level to characterise the profiles i.e. large movements at coarse resolution, small movements at fine resolution.

2.5.6 Local Feature Analysis

Penev and Attick [40] criticised the applicability of PCA on the grounds of its nontopographic (nearby values in the eigenmode index do not possess any relationship among each other) and nonlocal (the support of the kernels extend over the entire image) representation, and presented the Local Feature Analysis (LFA) paradigm for deriving local topographic representations for any class of objects.

The LFA representations are sparse-distributed and, hence, are effectively low-dimensional and retain all the advantages of the compact representations of the PCA. But unlike the global eigenmodes, they give a description of objects in terms of statistically derived local features and their positions.

A set of topographic kernels is constructed by means of the eigenmodes of the correlation matrix of an ensemble of images. Output correlation can be minimised by means of whitening and any orthogonal transformation of the eigenmodes (useful to produce representations meeting desirable properties). The receptive fields happen to be mostly local, and the outputs in a local region are correlated. The one that best describes it can be chosen to remain active and the rest suppressed. Thus each output predicts a small neighbourhood to an extent governed by the support of the output correlation. One possible strategy for sparsification is to represent the output with only a small subset of values so that the supports of the predictors cover the image space reasonably well.

Comparison with PCA and image subsampling show that the best perceptual reconstruction (least identity information of the error) corresponds to LFA, although the mean square error is lower with PCA.

2.5.7 Facial deformations

Another criticism about eigenfaces is its physical unfeasibility: images are modelled by a linear superposition of modes of intensity variation, without keeping any point correspondence whatsoever between any two faces, other than the alignment of the eyes. Although still a valid mathematical means for representing faces, the overlaying of intensity values is clearly a nonlinear process which results in apparent artifacts when the shape of the probe face significantly differs from the average face.

In [70] faces are modelled as deformable intensity surfaces, and a method for learning physical deformations is proposed. The intensity surface of the image is modelled as a deformable 3D mesh in the \((x, y, I(x, y))\) space. A set of representative deformations within a class of objects (e.g. faces) are statistically learned through a Principal Component Analysis, thus providing a priori knowledge about object-specific deformations. The approach dramatically reduces the computational cost
of solving the governing equation for the physically based system by approximately three orders of magnitude.

The image is represented by a mesh of $n \times n'$ nodes. The motion equation follows a Lagrangian formulation with mass, damping and stiffness matrices as well as an external (image) force. The unknown motion vectors are first projected into a modal subsbasis given by the most significative motion modes satisfying the motion equation. After that, the resulting vectors undergo a Principal Component Analysis and the vectors are projected in a reduced space. Hence, the resulting displacement is constrained to lie along those learned deformation modes that are characteristic of the object class (frontal views of human faces in this case).

2.5.8 Linear discriminant techniques

In [39] it is claimed that a successful automatic face identification system should be capable of suppresing the effect of factors such as changes in expression, 3D orientation, lighting conditions, hairstyles and so on, allowing any face to be rendered expression-free with standardised 3D orientation and lighting.

Shape variation reflects both within-class and inter-class variation. Isolated inter-class variation can be computed using discriminant analysis techniques. Discriminant analysis techniques are employed to ensure that within-class variation parameters are suppressed in favor of those controlling inter-class variation. Only 6 discriminant variables were needed to explain most of the inter-class variation. The discriminant parameters deform the model only to generate shapes representing different individuals. No changes in 3D orientation or expression are activated so that a system independent of expression and viewpoint can be developed. For identification, the facial features are localised using ASM [67]. Model points are transformed into shape model parameters and the face is deformed to the mean shape. Class (person) assignment is made according to Mahalanobis distance (the multivariate distributions are estimated during training). The application of discriminant analysis is also reported in [71].

In accordance with [39], an interesting point is made more recently in [72]: the variations between the images of the same face due to illumination and viewing direction are almost always larger than image variations due to change in face identity. In this work, 4 methods are compared:

- Correlation: all the images are normalised to have zero mean and unit variance to make the results independent of light source intensity. Problems arise when training and test images are gathered under varying lighting conditions.

- Eigenfaces: by throwing out the first principal components, the variation due to lighting is reduced, but at the price of losing useful information for discrimination, since it is unlikely that the first several principal components correspond solely to variation in lighting.
2.5. Descriptors and matching techniques

- **Linear subspace**: given three images of a Lambertian surface from the same viewpoint taken under three known, linearly independent light source directions, a 3D basis can be constructed for the linear subspace. Some of the problems arisen are due to self-shadowing, specularities and the fact that facial expressions have variability that does not agree with the linear subspace model.

- **Fisherfaces**: find projection onto the linear subspace above that maximises the between-class scatter to within-class scatter ratio.

It is concluded that Fisherfaces appear to be the best method to deal with variations in lighting and expression at the same time. The linear subspace method suffers when confronted with variations in facial expression. The images no longer lie in a linear space whilst the resulting projections following Fisher’s method tend to mask the regions of the face that are highly variable.

More recently, successful application of Linear Discriminant Analysis techniques to a face verification task on the XM2VTS database has been reported in [73].

### 2.5.9 Probabilistic matching

Moghaddam et al [74] argue in favor of a probabilistic measure of similarity, in contrast to simpler methods which are based on standard L2 norms (e.g., template matching) or subspace-restricted norms (e.g., eigenspace matching). The proposed similarity measure is based on a Bayesian analysis of image differences. Two mutually exclusive classes of variation between two facial images are modelled: intra-personal (variations in appearance of the same individual, due to different expressions or lighting) and extra-personal (variations in appearance due to a difference in identity). The high dimensional probability density functions for each respective class are then obtained from training data using an eigenspace density estimation technique and subsequently used to compute a similarity measure based on the a posteriori probability of membership in the intra-personal class, which is used to rank matches in the database.

Differences of eigenface coefficients are used to construct an intra-personal variation subspace and another extra-personal variation subspace. Although the described method does not address shape or point correspondence, so that it is arguable the ‘meaning’ of intensity-based vector differences, the metric should be better than direct comparison of the eigenface coefficients. It would also be interesting to determine to what extent the proposed subspaces are really representative for other databases. For instance, the intra-personal subspace is reported to form a dense cluster, which is not that obvious under general image conditions (despite photometric normalisation) in the light of the observations made in [72] about intra-class variability.

\(^3\)In fact, this algorithm was found to be the top performer of the ARPA’s 1996 FERET face recognition competition [75].
2.5.10 Active Appearance Models

Previous approaches to model fitting have used partial models, such as shape only or texture only; fitting full photo-realistic models has been limited to simple euclidean transformations, or else has involved complex interleaving of shape and texture fitting. The AAM [42] contains a statistical model of the shape and grey-level appearance of faces, and can generalise to most valid examples. The relationship between grey-scale differences and AAM parameters is found through linear regression operated on a training set.

The complete description of the face provided by the AAM is used as a basis for face recognition [76], and it is shown how the AAM framework allows identity information to be decoupled from other sources of variation (although it is admitted that the within-class spread takes a different shape for each different identity), allowing evidence of identity to be integrated over a sequence. The identity parameters consist of the projections onto an ‘identity’ subspace found using Linear Discriminant Analysis plus a class-specific linear correction. Tracking can operate at low resolution as evidence is integrated over the sequence.

2.5.11 Hidden Markov Models

Hidden Markov Models (HMM), well known for their application to speech recognition [77], have also been used for face recognition. In [78] it is shown how these models allow for the automatic extraction of facial features and the classification of face images and some experiments are presented to support the plausibility of the approach. An 8-state ergodic HMM and a left-to-right model were compared on a 20-subject database constrained to homogenous lighting and constant background. The results were considered successful, with a better performance in the case of the left-to-right model.

In [79] and [80], further details can be found about the use of left-to-right HMMs for face recognition. Frontal images of faces are sampled using top-bottom scanning following a raster pattern. As a result, there is a natural order in which the features appear and this can be conveniently modelled using a top-bottom HMM (left-to-right in the literature, see e.g. [77]). Three HMM parameters affect the performance of the model: the number of states $N$, the height of the sampling window $L$ and the amount of overlap $M$.

2.5.12 Neural networks

The use of neural networks [81] for face recognition has also been investigated. An example of this model-less approach is reported in [82]. Images of the user at four different scalings are used as correlation templates to adjust scale and translation variations. For enrollment, users are cautioned to present their images squarely to the camera (slight turns left, right, up and down are allowed). Also unacceptable images (large head turns, occlusions, varying expressions) are collected as counterexamples.
The neural networks are standard feed-forward perceptrons with a single hidden layer. Training sets consist of user's pictures as exemplars and all other users' pictures as counterexamples. One network uses as input a 1400 element vector obtained by scanning the test image. The other uses the normalised projection of the test image with a set of eigenfaces.

Another example where neural networks are used is described in [83]. Grey scale projections are used as features for verification and recognition under the claim that a small number of projections can provide sufficient information for classification and are relatively immune (horizontal projections) to rotations about the vertical axis. The working database consisted of 55 subjects (9 images each, encompassing different expressions and rotations plus different scale). It was found that transform coding (Discrete Cosine Transform) of the projection waveform minimised the effect of local changes due to expression. Further adjustments for changes in scale can be accommodated using dynamic time warping. The proper face recognition system uses a hierarchical pattern classifier, the Neural Tree Network (NTN), that combines neural networks and decision trees. The NTN grows as it learns, without the need to specify the number of neurons in advance.

As mentioned in [17] neural network methods can potentially incorporate both numerical and structural information, and their ability to generalise and recognise using incomplete information is pointed out, but their usefulness needs to be evaluated on significantly larger and more representative databases. How to select an appropriate architecture is still a crucial research issue (typically resorting to arbitrary heuristics), and so is the need for sometimes interminable training, which sounds a bit shocking when contrasted with the functional characteristics of the human nervous system, which can learn a pattern from a substantially lower number of training samples.

2.6 Addressing face variability as a 3D object

Many of the problems associated with face recognition are due to its intrinsic 3D character, whilst most of the approaches described in the literature are view-based and analyse the face as a 2D object, although aware of its authentic 3D nature. Alternative approaches have been directed towards the recovery of the real 3D structure. The most common of these approaches use some sort of structured light pattern [84, 85, 86, 87] to extract volume information through the distortion inflicted on the patterns as projected on the face. The use of range data has not received the same attention, mainly because of its high cost for commercial applications. Another currently used approach is based on multiple 2D views. Although it is not properly a 3D method, the variability associated with the characteristics of a 3D surface can be captured up to a certain extent.

2.6.1 Varying pose

As pointed out in [88], while some systems, especially template-based ones, have been quite successful on expressionless, frontal views of faces with controlled lighting,
not much work has taken face recognition systems beyond these narrow imaging conditions.

In order to build a face recognition system that works under varying pose, the difficulty of it is to handle face rotations in depth. Building on successful template-based systems, an approach is presented in [88] to represent faces with templates from multiple model views that cover different poses from the viewing sphere. To recognise a novel view, the recogniser locates the eyes and nose features, uses these locations to geometrically register the input with model views, and then uses correlation on model templates to find the best match in the database of people.

In a further work [89], pose-invariant face recognition from a single view is addressed. Given one real view at a known pose, it is still possible to use the view-based approach by exploiting prior knowledge of faces to generate virtual views, or views of the face as seen from different poses. To represent prior knowledge, 2D example views of prototype faces under different rotations are used. Example-based techniques are used for applying the rotation seen in the prototypes to essentially ‘rotate’ the single real view which is available. Next, the combined set of one real and multiple virtual views is used as example views for a view-based, pose-invariant face recogniser.

Face representation is divided into shape (interocular-distance normalised locations of a collection of feature points) and textural (original gray level image warped to a shape-free representation) components. The experiments suggest that among the techniques for expressing prior knowledge of faces, 2D example-based approaches should be considered alongside the more standard 3D modelling techniques. A related work, namely image synthesis from an example view, can be found in [90].

### 2.6.2 Profiles vs frontal views

In [91] a frontal face recognition system based on second order statistics (SOS) and a profile recognition system based on P-Fourier descriptors (PFD) are compared. PFDs can be applied to open and closed curves, are invariant to parallel transformations and enlargement / reduction, and have simple relation to rotation. SOS consists of correlation computed on edge images. The conclusions drawn from this comparison test are that frontal faces include more information than profiles, but are more sensitive to tilts than profiles, as shown in other experiments performed. It happens in the opposite way when left/right rotation views are considered, where profile based recognition performs poorly.

Taking into account that the information supplied by each of these modalities is expected to be only partially correlated the combination of both has also been investigated. This approach is followed by Pigeon and Vandendorpe [92], who describe a system where a profile-based expert is combined with a frontal-based expert, resulting in a more robust image-based person authentication system which is measured on the M2VTS multimodal face database [57].
2.7 Discussion

Daugman [21] mentions that the face recognition performance of humans has an error rate slightly below one per cent (no indications is provided about the context in which this rate is supposedly obtained) and questions whether algorithms can be designed that might approximate that figure. He also refers to another upper bound for face recognition set by the birth rate of identical twins, which also happens to be slightly below one per cent. This test does not seem to have been passed so far by any computer vision algorithm. Humans, however, are normally able to distinguish between identical twins who are well known to them by picking up on any phenotypic differences.

On the other hand, humans are not so reliable at face recognition as we might believe. A well known example [93] among the face recognition community was provided by Sinha and Poggio, who presented a photograph showing two famous characters standing next to each other where the faces had actually been swapped, fooling the human visual system into mistaking their identities. According to Sinha and Poggio, the illusion would be a result of brain 'shortcuts', i.e. to recognise a face rapidly, time after time, the brain would home in on certain salient facial attributes, such as head shape, namely hair style.

Testing the actual performance of face recognition algorithms is a thorny issue. Most of the reported results lack statistical significance because of the limited size of the test databases, and because of the representativity of the images contained therein. The Face Recognition Technology (FERET) program [75] addressed the issue of creating a large database of facial images (more than 14,000 images from more than 1,100 individuals) and the testing procedure to evaluate systems. Despite the a priori large number of different individuals and images, the FERET database is obviously not entirely representative of facial appearance variability and the results reported by the researchers having taken the tests should be carefully analysed, both in the light of the limitations of the database and the characteristics of the algorithms being tested.

Despite the undeniable improvements in facial image understanding, there still subsists the problem of finding a realistic model of human faces, general enough as to capture the essence representing any human face, and specific enough in order to be able to characterise any single face in an unambiguous way.

Anyhow, even at the 99% level, assuming this limit is achieved in realistic test scenarios, would not be that good for high-security applications\(^4\). However, despite the shortcomings of face recognition as a solid candidate for high confidence applications, its importance in daily human interaction is indisputable and, quite in the way humans seem to integrate information from different sources to form a certain hypothesis, more reliable recognition engines can be built by fusing the information

---

\(^4\) With regard to high reliability biometrics, the random texture that is visible in the iris of the eye has gained considerable credit for recognition of personal identity with very high confidence levels. The randomness of the iris pattern allows for a significant number of meaningful degrees of freedom, and also presents very interesting properties, such as its stability over life and its distinctiveness (unexisting correlation between monozygotic twins, and between the two eyes of a given person).
supplied by a number of recognition modalities, none of them inherently reliable, as was the scope of the M2VTS project [2].
Chapter 3

Lip-reading

3.1 Introduction

Lip-reading\(^1\) is the name given to the process of extracting relevant image-based information from lip motion based on the way hearing-impaired people manage to do speech recognition. However, this modality is not exclusive to hearing-impaired people and studies have shown that visual information from a speaker's face is integrated with auditory information (e.g. McGurk effect [94]), as will be seen in more detail in section 3.2.

Other researchers [95] have also reported how the performance of speech recognition systems might be improved if both acoustic and visual inputs are combined. Hennecke [96] remarks the complementarity of the information provided by the talker's image to enable improved recognition accuracy, especially in environments corrupted by high acoustic noise or multiple talkers. Because most of the phonologically relevant visual information is from the mouth and lips, inferring their dynamics in an accurate and robust way is claimed to be very important. A similar claim regarding the relevance of lip dynamics is made in [97], where it is stated that real information in lip-reading lies in the temporal change of lip positions rather than in the shape.

Cohen [98] deals with the other side of the coin: 'visual' speech synthesis. Again, emphasis is put on the robustness of human speech recognition performance because of multiple sources of information used, and comments are made about the usefulness of speech synthesis to test speech recognition systems. With the development of multimedia applications and model-based coding, there has been an obvious interest in determining which parameters are of relevance for representing mouth shapes, namely for coding, or for driving talking faces or any other kind of synthetic personae [99, 100].

The improvements in speech recognition performance achieved by combining both acoustic and visual information motivated the consideration of lip-reading for speaker verification within a multimodal framework, which was the objective of the M2VTS

\(^1\)It is common in the literature to find references to the rather speech-focused *speechreading* term as a synonym.
project [2]. Lip-reading techniques (e.g. lip trackers are particularly interesting in as much as they allow the automatic characterisation of lips over a sequence of frames) used for speech recognition are also very useful in a speaker recognition/verification scenario but the focus will obviously be on extracting those characteristics which are specific of a particular speaker. Consequently a survey on automatic lip-reading for speech recognition applications is presented in section 3.3, whereas section 3.4 deals more specifically with lip modelling and tracking. Dynamic contours, a particularly interesting contour-based approach, will be described in section 3.5.

3.2 Lip-reading by humans

Visual speech cues play an important role in the acquisition of speech perception and speech production from early life on. As reported by Liittin [101], children are aware of the existing congruence between lip movements and speech sounds as early as three months of age, and speechreading skills are acquired at early age as well: toddlers can speechread familiar words at 19 months of age. Visual speech also influences the process of learning to speak: blind children are slower in the acquisition of speech production for sounds which have visible articulation.

The complementarity of visual cues and the acoustic signal for speech recognition is well established [96]: some phonemes which are difficult to understand acoustically are easier to distinguish visually and the other way round. For instance, "ma" and "na" are highly confusable acoustically but easy to distinguish visually (different lip closure); "ba" and "pa" are easy to distinguish acoustically (different voice onset time - delay between the burst sound and the movement of the vocal folds) but highly confusable visually.

It is well known that not only hearing-impaired persons do make use of those visual cues for speech perception², but how humans integrate visual and acoustic information is not well understood. The influence of visual articulation on human perception of speech is demonstrated by the McGurk effect [94]. Subjects are presented simultaneously with an acoustic recording of an utterance and a visual recording of a speaking face corresponding to a different utterance. However, what the subjects perceive is a sound which is neither of the presented stimuli. For instance, a subject hearing a recording of /ba/ but seeing an utterance of /da/ results in perceiving /g/². Not all utterances result in such nice McGurk pairs, but this perceptual illusion is conclusive about the bimodality of speech. Further McGurk effects and the influence of both congruent and incongruent visual stimuli are analysed in [103]. It is reported how a visual speech signal needs to be particularly strong to impair identification of incongruent auditory speech whereas a weak visual speech signal is sufficient to improve identification of congruent auditory speech.

There are two main models attempting to explain the integration by humans of both visual and acoustical information according to how much processing is applied to

²The extraction of visual cues to help speech comprehension in particularly noisy environments such as a loud cocktail party -therefrom the so-called cocktail party effect- is a well established fact, as mentioned in [102].
3.3. Automatic lip-reading for speech recognition

the signals before they are brought together:

- Late integration: audio and video classifiers work separately on their respective information channels and yield a category for each utterance; both category responses would then be pooled to make an overall judgement. The model assumes that both data streams are conditionally independent. Temporal information across channels is lost, though.

- Early integration: the acoustic and the visual information is combined while still remaining relatively unprocessed; this mixed information is then jointly processed to provide the final categorization. The model assumes conditional dependence between both modalities and also accounts for temporal dependencies.

Stork [102] mentions the existence of a general consensus in the psychological community about humans' late integration for speechreading, the argument being that the information from both the eyes and the ears are processed in some detail by separate sets of neurons before the two streams can eventually meet and be integrated. However, this point remains controversial [101]: the auditory system seems to perform partial recognition which is independent across channels (different acoustic frequency bands) but the integration of acoustic and visual features appears to occur before speech is categorised phonetically.

3.3 Automatic lip-reading for speech recognition

This section covers several applications of lip-reading to speech recognition ranging from support for acoustically based speech processing to speaker verification.

In [104] an image processing system which can extract the velocity of the lips from image sequences to support speech segmentation is reported. The velocity of the lips is estimated by a combination of morphological image processing and block matching techniques. The resultant velocity of the lips is used to locate the syllable boundaries. The information is found particularly useful when the speech signal is corrupted by noise. The correlation between speech signals and lip information is also demonstrated and data fusion techniques are used to combine the acoustic and visual information for speech segmentation. Experimental results show that the combination of visual and acoustic signals can reduce segmentation errors by at least 10.4% when the signal-to-noise ratio is lower than 15 dB.

In [28] (see also the related [105] and [106]), a system is described performing face localisation and tracking plus lip-reading. A triple neural network approach is followed, first for face localisation and then for lip localisation and speech recognition. Sobel edge maps are used as the input of a neural network that learns the distribution of the Sobel edges around the mouth. It seems that lips can be localised in fairly low resolution images. Fine positioning is done by another neural net operating on horizontal edges. A Multi-State Time Delay Neural Network fed with parallel
acoustic speech and mouth images performs the recognition. The following lip descriptors are used: low resolution (24x16 pixels) intensity images, band-pass Fourier magnitude coefficients, Principal Components of the downsampled image and Linear Discriminant Analysis (LDA) of the downsampled image. The best recognition improvements (tested for different signal to noise ratios) by use of visual information corresponded to the LDA coefficients followed by the downsampled intensity image.

Alternative connectionist approaches combining acoustic and visual informations include Hidden Markov Models ([107, 108]) and hybrid architectures such as the hybrid MLP/HMM speech recognition system described by Bregler and Konig [97].

A visual speech recognition experiment is reported in [109] with only a 4% error rate, although the tests conditions were extremely simple: very small vocabulary (4 words), speaker dependent and it does not deal with continuous or spontaneous speech, but with isolated words. Similarly, a word accuracy of roughly 80% is reported in [107] using the Tulips 1 database [108]. The experimental framework consisted of gray-level images of 12 speakers uttering the first four digits of the English language. Two sequences were available (one for training, the other for testing). Tests were done following the leave-one-out strategy because of the small size of the database.

Similar tests on the Tulips 1 database involving left-to-right models HMMs are described in [108]. The tests were performed varying the number of states and mixtures per state following several preprocessing strategies. The tests were done following the leave-one-out strategy for 12 people x 4 digits x 2 sequences. A top performance of about 90% was achieved. The same tests done by trained hearing impaired persons yielded a correct recognition rate of 95.5%.

In general, it seems that visual information contributes to improve the recognition rate of stand-alone acoustic based speech recognition systems and that the improvements are related to the degree of corruption of the acoustic signal. Similar conclusions are also drawn by the AMIBE man-machine communication project. The project AMIBE (Applications Multimodales pour Interfaces et Bornes Evoluees) [12] addresses basically two aspects: oral message recognition (synchronising lip-reading with acoustic speech recognition) and user identity verification (check speaker voice model against encoded references). Physical parameters such as lip width, height and area are considered as well as their first and second temporal derivatives for the message recognition task.

In the framework of the M2VTS project [2], Lüttin [101] also reports improvements in both speech recognition and speaker verification when the acoustical information is augmented with visual features obtained by a lip tracking system.

### 3.4 Lip modelling and tracking

Some of the earliest lip models [95] consisted of mouth shape codebooks constructed by means of clustering techniques applied to thresholded gray-level regions. The codebooks could then be used as templates for comparing the visual characteristics
of different utterances. Thereafter, a variety of lip models can be found in the literature, ranging from model-less image templates to rather complex 3D models.

### 3.4.1 Deformable templates

The so-called deformable templates [50] are parametric, geometric models of an object. Variations in the parameters correspond to allowable deformations of the object and can be specified by a probabilistic model. After the parameter extraction stage, the parameters of the template can be used for object description and recognition.

Hennecke [96] describes the use of deformable templates for speechreading, in order to infer the dynamics of lip contours throughout an image sequence. The template computations can be done relatively quickly and the resulting small number of shape description parameters are quite robust to visual noise and variations in illumination. Such templates delineate the inside of the mouth, so that the teeth and the tongue could also be found. The templates are computed from gray-level images manually cropped around the mouth without the use of invasive markers or patterned illumination. The template consists of parabolas and quartics which follow the outer and inner edges of the upper and lower lips.

The tracking of lip heights is usually satisfactory and stable, but determining widths proves much harder. Some of the problems are attributed to speaker variability and lighting conditions. Moreover, the chin and teeth could provide distractors. A similar model for the lips consisting of 4 parabolas is described in [110], also for speech recognition. Problems are reported when the lip estimates are attracted to an improper edge or to two at the same time.

Coianiz [14] follows the same model described in [96] but uses colour instead to determine optimal values for the parameters of a 2D deformable model of the mouth. The mouth vertices are localised using luminance taking advantage of the fact that the inner profile of the lips is visually well distinct. Apices locations are found using chrominance.

### 3.4.2 Mouth region descriptors

Bregler [97, 111, 109] describes an eigenrepresentation of the mouth region as a whole. Snakes are firstly used to localise the lip contour so as to accurately center gray level images around the mouth. The contours are coded as 80-dimensional vectors by concatenating the coordinates of 40 image points. Initially, this set of configurations was projected onto a 5-dimensional space following PCA, but it was found that the outer boundary of the lips was not distinctive enough to give reasonable recognition performance. On the other hand, the inner lip contour and the appearance of teeth and tongue were deemed important for recognition. As a result, the contour-based lip tracker is only used to center a grey-level matrix from which the features used for speech recognition are extracted.
A 5-dimensional manifold is obtained by a procedure that performs K-means clustering followed by principle components analysis on points in the neighbourhood of the cluster centers. The projection of a point onto the manifold (ie. the nearest point on the manifold) is achieved by projection into all tangent planes and then finding a weighted mean. The weighting is done by Gaussians with variances determined by the local sample density. Lip tracking is done by point-wise gradient descent followed by a projection onto the 5-dimensional manifold.

3.4.3 Contour-based approaches

Contrary to [109], the lip contours are claimed to be distinctive enough and to convey relevant information to support speech recognition. In [15, 107] an Active Shape Model (ASM, see [67]) is used for lip modelling. The lips are represented by a set of labelled points along their contour.

The ASM training set is labelled by hand and the mouth width is used as a reference length to map different examples of lip shapes. The points are evenly spaced along the horizontal direction, so that only vertical parameters are required to describe a shape.

Image gradients are found inappropriate for representing lip boundaries on the grounds of great variability along the contour and dependence on speaker, illumination, reflections, tooth visibility and mouth opening. Profile information is used instead. 1-dimensional profiles are extracted perpendicularly to the contour at each labelled point location and all the profiles are then concatenated into a global profile vector on which PCA is performed. Thus, a model is built that describes the mean intensity profile of the training set and its main modes of variation.

The parameters describing the shape of the lips are extracted and used as visual speech feature vectors of a Hidden Markov Model used to model the temporal changes. Each utterance is represented as a sequence of visual speech vectors. The emission probabilities are modelled by continuous Gaussian distributions. Problems have been reported with the inner lip contour. The profiles inside the inner lip contour change mainly between three different levels, so that it is unlikely that a Gaussian distribution is appropriate to describe the variation.

The so-called dynamic contours or active contours e.g. [112, 113] tackle tracking as a two-phase process in which a dynamical model is used for prediction -to extrapolate motion from one discrete time to the next- and then the prediction is refined using measured image features. Section 3.5 specifically deals with these methods.

3.4.4 Other approaches

In [108] lip images are modeled as mixtures of independent Gaussian distributions, and the temporal dependencies are captured with standard left-to-right hidden Markov models. The results indicate that simple hidden Markov models may be used to successfully recognize relatively unprocessed image sequences. The system
achieved performance levels equivalent to untrained humans when asked to recognize the first four English digits (Tulips 1 database, UCSD, USA).

Matthews et al [114] use features derived from a one-dimensional multiscale spatial analysis (MSA) of the mouth region using a new procedure termed the 'sieve' [115] derived from mathematical morphology and median filtering to characterise mouth shapes. The sieve uses morphological filters to simplify signal representation over multiple scales, preserves scale-space causality and can reversibly transform a signal to a granularity domain. As an alternative to wavelet decomposition, a scale histogram computed from the grey scale pixels of an image containing the mouth can be used to characterise this area on a per frame basis. This is independent of image amplitude and position information and neither accurate tracking or special markers are required.

Blob features, i.e. spatially-compact clusters of pixels that are similar in terms of low-level image properties, are described in [116] for representing mouth shapes.

Basu et al describe a 3D model for the tracking of human lip motions in [117, 118]. First a physically-based 3D model of lips is built and trained to cover only the subspace of lip motions. The model is tracked in a video sequence by finding the shape within the subspace that maximizes the posterior probability of the model given the observed features. The features are the likelihoods of the lip and non-lip colour classes. Forces are iteratively derived from these values and then applied to the physical model so as to converge to the final solution.

### 3.5 Dynamic contours for lip tracking

A classical approach for tracking of contours is the well known snakes method [52]. A snake is an energy-minimising spline guided by external constraint forces and influenced by image forces that pull it toward features such as lines and edges. Snakes are active contour models: they lock onto nearby edges, localising them accurately. The snakes rely on other mechanisms to place them somewhere near the desired contour. Scale-space continuation can be used to enlarge the capture region surrounding a feature. Snakes applications in visual problems include detection of edges, lines and subjective contours, motion tracking, and stereo matching. An example of the use of snakes to construct and edge-based lip tracker is already reported in this paper. The spline is initialised to the speaker's lips in the first frame and then automatically tracks the lip movements.

Later works have included new modifications or extensions to the original dynamic contour approach based on the minimisation of an energy along the tracked contour. For instance, in [119] new active contour models can be derived by multiplying the Euclidean arc-length by a function tailored to the features of interest to which we want to flow. This amounts to defining a new Riemannian metric. The new metric has the property that it becomes small where the weighting function is small and vice versa. At such points lengths decrease and less 'energy' is required in order to move.
Chapter 3. Lip-reading

In [120] a simplicial grid (triangle grid for 2D spaces) is used to iteratively reparameterize the model, allowing it to flow into complex shapes and change its topology when necessary. The use of an inflation force for the snake model eliminates the need for an inertial force term. Simplicial decompositions provide an unambiguous framework for the creation of local polygonal approximations of a contour or surface model. The reparameterization can be performed every iteration or every nth iteration to improve computational efficiency. The grid can be turned off at any time, reverting to the traditional snake formulation. For automatic segmentation, an approach is proposed consisting of the distribution of a set of small circular snake 'seeds' over the image domain. If there is no object of interest to attract a snake, it will shrink and eventually disappear.

B-splines [121, 122, 123] have received considerable attention for tracking, mainly because of their lower computational cost and easy parameterisation, since a small subset of control points is enough to fully represent a B-spline contour. Thus, the number of variables to be estimated is reduced to the number of control points. Other properties of B-splines include local control and continuity.

B-spline approaches in this section have been grouped according to the measurement model used for feature search. Earlier works perform B-spline updates during tracking based on an assumed motion model according to the current feature position estimate, which is supposed to correspond to the ground truth feature. Later works follow a statistical approach where the sensor characteristics and the measurement uncertainty are taken into consideration. The application of Kalman filters [124, 125] is commonly referred to in these approaches.

3.5.1 Deterministic physically-based dynamic models

In [126] B-splines are used to reduce the state space of the problem. The dynamics model includes internal forces as well as external forces. The internal forces enforce smoothness, whilst the external forces guide the active contour towards the image feature (coarse to fine search). The image gradient is used as the guiding information. Emphasis is made on how appropriate dynamics modelling can greatly enhance tracking performance. The physical properties modelled include inertia, attraction towards feature and velocity damping. The tracker follows a parallel implementation using a network of transputers to favour real time processing. Tracking results show steady state errors for constant-velocity moving objects, in accordance with the response of a 2nd order system to a ramp input [127].

Some extensions to this work are found in [128], where coupled contours are used to track objects at video rates. Contour coupling involves the incorporation of a strong tendency to a particular relaxed configuration in order to reduce the tendency of the dynamic contour to be distracted. The spline contour is coupled to a rigid B-spline template. The template is learnt by allowing an unconstrained dynamic contour to lock onto a representative target feature. Tracking dynamics follows a Lagrangian formulation of the contour motion. Contour motion can be decomposed into modes (eigenstates). The tuning of each modal control system is different. However, tuning is interdependent across models and this limits tracking performance.
3.5. Dynamic contours for lip tracking

The assumption underlying coupling is that feature shape changes very slowly. Coupling is defined by an elastic energy term in the Lagrangian equation. A complete modal analysis is only done for single-span snakes. Rigid translation corresponds to the first mode of variation. Not all the modes can simultaneously be critically damped. Higher modes than one will be overdamped, which means they will be slower than ideally desired. Under steady state motion, the translational mode showed a constant lag, but higher modes exhibit a lag that increases over time due to the influence of the template, which leads to distortion effects.

The time step used to discretise the system must be small with respect to the highest frequency component of the expected solution. Implicit differencing is used to retain computation speed. However, this implicit Euler scheme does not exactly solve motion equation; rather it shows additional damping and a lower effective natural frequency.

The implementation exploits a parallel architecture using a network of transputers. The system will successfully track features whose velocity is such that the lag caused by viscous drag does not exceed the radius of the tracking window.

3.5.2 B-spline tracking with Kalman filters

In [112] a mechanism is developed for incorporating a shape-template into a contour tracker via an affine invariant coupling. In that way the tracker becomes selective for shape and therefore able to ignore background clutter. Affine invariance ensures that the effect of varying viewpoint is accommodated. The use of a Kalman filter allows uncertainties to be treated systematically, which accommodates object flexibility and unmodelled distortions such as the deformation of a silhouette under motion.

Feature search occurs on a specified spatial scale defined by a search window. The diagonal gain matrix of the Kalman filter implies that the second order dynamics in state space degenerates into a set of identical, independent 2nd order systems. There is an automatic control of the spatiotemporal scale. In the absence of a feature, the search scale grows as $t^{3/2}$. As the whole feature locks on again, the search scale will contract again as $\frac{1}{t}$ until it reaches the steady state scale. Analogously, the temporal scale varies as the inverse square of the spatial scale. In the absence of a feature, temporal scale shortens so that feature acquisition and locking can occur rapidly. It is pointed out as a crucial factor that a shape-specific tracker should continue and retain some shape memory. A mechanism is provided to extinguish those components of shape that are not related by affine transformation to the template used to stabilise the spline shape.

Quadratic B-splines with the possibility of multiple knots for vertices are used for curve representation in [129]. A 'hand-drawn' template is used to stabilise the spline shape. Further to that, the spline is projected into a state-space of reduced dimensionality constructed as the space of allowed deformations of the curve relative to the template. The motion model consists of a second order differential equation driven by noise. The measurement process involves casting rays along several normals to the estimated curve. The feature is a high contrast edge searched for within a $\pm 40$
pixel window. The measurement procedure has the effect of pulling the tracked curve along the normal only. Tracking involves a prior training process: a default simplified tracker is used to gather a sequence of images with which the parameters of the trained tracker are found. This process is done off-line. The typical number of measurements per B-spline span is set to 3 whereas the number of control points is of about 20. The resulting tracker is specifically sensitive to translation, but non-rigid motion can also be dealt with by refining the measurement space. This is done by means of the so-called key-frames: typical nonrigid deformations on which the tracker contour is positioned interactively. The template and the key-frames do not need to be mutually orthogonal, merely linearly independent.

In a related work [130], emphasis is made on the need to limit the number of free parameters (apparently conflicting with the requirement of a high number of control points for accurate shape representation) to ensure tracking stability. Analogously, in order to infer system dynamics, a training set is obtained using a default tracker based on constant velocity and isotropic plant noise. Examples of tracking are available in configuration space (concerned about shape tracking, as the static component of the object model) and in phase-space (motion characteristics).

Analogously, tracking through motion learning is reported in [13], where image sequences are used to learn parameters in a stochastic differential equation model. The disadvantage is that each tuned tracker is effective only for the relatively narrow class of shapes and motions on which it was trained. Again, quadratic B-splines are used to represent the contours. Tracking bootstrapping is done using a default (simplified) tracker.

In the case of nonrigid motion, the extra degrees of freedom are derived from key frames as described above. For both rigid and nonrigid motions, tracking performance can be greatly enhanced by training. One of the example deals with lip tracking (side view against an uncluttered background, to avoid the use of any highlighter). The default tracker is used to follow a slow-speech training sequence and thus build an improved tracker, which is used to follow a medium-speed sequence and train the final tracker, used to perform under real conditions. A full Kalman filter with time-varying gains and validation gains enhances the tracker's power of recovery from loss of lock.

In general, contours cannot be assumed to be slowly varying. This is the point made in [131], where it is claimed that the tracking method must react well to large shape deformations. Training shapes are represented by closed uniform B-splines with a fixed number of control points equally spaced around the boundary. The curve is sampled at \( m \) regular intervals between control points. Three elements can be distinguished: translation, shape alignment (rotation and scaling), and shape variability. The change in origin is estimated using a dynamic Kalman filter; then the change in alignment parameters is computed and, finally, the shape parameters are estimated using a Kalman filter. It is also proposed, alternatively, to update simultaneously all the parameters if the rotation, scaling and translation effects are sufficiently small. In order to ensure that the shape contour is feasible, the shape parameters are constrained to lie within a hyperellipsoid. The imposed global constraints stabilise the system and prevent tangling. Once the state estimates are
3.5. **Dynamic contours for lip tracking**

updated, the contour may still be inaccurate (for instance, if the change between frames is too large for the filtering process). An iterative procedure can be used to refine the estimates, but it can also lead to a situation where the state covariance matrix values become unrealistically small. The $\epsilon$ method [124] is used to avoid such situation.

In a related work [132] the control points of a cubic B-spline are treated as landmark points to generate linear Point Distribution Models [67]. Each shape is represented by a shape vector consisting of nodal parameters that represent an approximation to the curve. The boundary can then be approximated using a cubic B-spline with a fixed number of control points. An active search mechanism is implemented for tracking. For the first frame of the sequence, the estimated variance of each shape parameter is initialised to zero. Noise terms are added in subsequent frames to allow the shape parameters to vary slowly. It is assumed that high quality training images are available in which the approximate location, size and orientation of the object are known. Iterative variations which do not occur within the initial training set will never become apparent in subsequent models. The current eigenshape model is perturbed by a simulated noise process to allow fine detail that is not well represented in the original model. The Kalman filter allows the more significant modes to vary more easily. The model has been successfully applied to a problem of tracking the outline of a walking pedestrian in real time.

### 3.5.3 Refinements and extensions

A novel stochastic algorithm, CONDENSATION, for Conditional Density Propagation over time, was presented in [133]. The system is proposed to deal with object tracking in cluttered backgrounds, where Kalman filter-based systems work relatively poorly, since the clutter easily 'distracts' the spatio-temporal estimate. Specific probability densities must be established both for the dynamics of the object and for the measurement process. Second order dynamics models are used whose coefficients are learned from sequences of images. An untrained tracker is used first to follow training motions against a relatively clutter-free background. Examples are available of tracking multi-modal distributions (walking person in a room with other people in the background), tracking of rapid motions through clutter (the Kalman filter was distracted irrecoverably after switching to the clutter peak because of its higher *a posteriori* probability at a given time during tracking) and tracking of complex jointed objects (flexing hand moving across a cluttered desk). Notwithstanding the use of stochastic methods, the algorithm is claimed to run in near real time.

Another work [134] addresses the issue of background modelling. Edge-detection for feature search is replaced by a statistical test to determine if each point on each searchline is foreground (object) or background. Approximate camera calibration is sufficient since any residual error will be absorbed into the statistical model. Single Gaussian distributions do not represent well the underlying distributions near high contrast edges whereas a two-Gaussian mixture would appear to be adequate. The single Gaussian model is also deemed inadequate when the foreground interacts with the background (e.g. casting shadows). Correctly fitting a two-Gaussian model to
points in the image takes a long time, but only a small amount of them would need a two-Gaussian mixture to represent their intensity distribution (their variance is checked first).

Two interesting aspects of tracking are covered in [135]. The first concerns modelling object classes for objects in motion. The second addresses the efficient modelling of coupling between tracked objects. Class variability can be learned by PCA applied to a linear curve parameterisation. Modelled class variability is only considered as the tracker is re-initialised on a new object. Once tracking is underway, the identity of the object does not change.

As far as object coupling is concerned, failure to track the horizontal translation of lips [45] is illustrated as a situation where tracking instabilities could be avoided by considering the motion of the object in which the tracked object is embedded. One object is primary and drives a secondary object. The motion of the primary object is assumed independent of the secondary. Coupling between the objects is established in terms of position and velocity. The computational cost is evaluated at three times the cost of tracking the two objects in an independent way. The computational requirements can be lowered down by means of a weakly coupled approximation that is constructed by treating the state of the primary system as if it were known exactly. An example of a tracker for the motion of head and lips is shown. The system becomes resilient to horizontal translation instabilities. The weak coupling produces smoother estimates of the lip shape than strong coupling and at a lower computational cost.

3.6 Discussion

In [101], the author comments on how human speechreading might be limited due to the temporal resolution of visual events\(^3\) and how a machine speechreading system could easily improve the temporal resolution and might recognise more visual events than humans. Analogously, subtle changes in facial expression which are hardly noticed by the human eye, can be successfully detected by a computer system, as reported by Rosenfeld in [102].

References to lip-reading applications for speech recognition and synthesis are abundant in the literature but, other than the work done by Lüttin [101] or Mason et al (e.g. [136]), there does not seem to be that much research on lip-based speaker verification (recognition). Visual features extracted from the mouth area rely on some sort of tracking and, as can be seen in Lüttin’s work, but also in the inspiring work by Blake [112, 113], it is common to limit strongly shape variation to ensure stable tracking performance. Whereas this looks all right in principle for (audiovisual) speech recognition, speaker verification may be seriously hampered by limitations that suppress detail that could be potentially useful to distinguish among mouth

\(^3\)It is estimated that humans can only distinguish about 8 to 10 lip movements per second - as opposed to an average of about thirteen different speech sounds per second in acoustic speech-, and some studies cited in [101] conclude that human speechreading skills depend highly on the speed of low-level visual neural processing.
shapes corresponding to different speakers. Lüttin addresses this point by deriving his eigenlips from roundabout 1000 images from the Tulips 1 database [108]. However this is done by manually annotating landmark points in the images, which is inconvenient and not particularly open to expandability should larger databases be considered. Therefore, there was a need for methods aimed at automatically determining modes of shape variation allowing for as much interpersonal variation as possible to support person identity discrimination. This will be the subject of the following chapters.
Chapter 4

The bootstrap tracker

4.1 Introduction

Tracking of lip contours has attracted the attention of the research community [95, 15, 137] because of the additional information conveyed to complement voice-based speech recognition techniques. Tracking of lip contours in side views against a constant background was successfully reported [129], but frontal views without artifacts such as lipstick had proved much more complicated and this precluded fully automatic lip tracking in the past. In the case of grey-level images, although some features can be quite consistent, such as the commonly referred lip intensity valley [45], standard derivative-based methods cannot be relied on and fail quite often to locate the lip boundaries in areas of particularly poor contrast, such as the lower lip. On the other hand, pure intensity-based methods can be expected to be very sensitive to lighting variations and shading. Successful results have been reported [15] using grey-scale point distribution models [67], but it is still unclear how to extract those models from the data in a reliable, automatic way.

Previous tracking experiments performed on the M2VTS database [57] exploited simple derivative-based cues. For instance, the outer lip contour was searched looking for dark to bright transitions, under the assumption that, in general, lips are darker than the skin in the mouth area. Even with good frontal illumination, in the absence of intruding shading, the above reasoning proved too weak and could only work when working close to the ground truth edge. However, even in the case where tracking penalised distance from the present feature location and the initial fit was good, tracking robustness was too weak and the tracking too sensitive to similar dark-bright transitions occurring in the neighbourhood. Other sources of problems or instabilities originated from specularities and poor contrast conditions, especially in the lower lip area.

Similar approaches using colour images were only slightly better. As a result, simple differencing-based algorithms were replaced by statistical methods. Chromaticity models were preferred because of their independence to intensity changes, as each colour component is normalised by the intensity value (obtained by averaging the 3 RGB components). In [138] it was shown how colour information can be exploited
Chapter 4. The bootstrap tracker

to perform a lip boundary tracking without hardly any supporting constraints other than the smoothness and the relatively reduced number of degrees of freedom of the B-spline representation (e.g. [121, 122, 123]) used to characterise the lip contour.

Colour characteristics clearly supply information that is absent in grey-scale images, the limitations of which are also reported in [139], and their success have justified the switch to a colour-based tracker. However, colour information in itself is not sufficient to robustly track the lips, as there are other areas in the human face with similar colour characteristics, and, anyway, colour measurements can be particularly noisy or misleading at some times as to rely too much on them.

It is also well established that the application of shape and/or dynamics constraints (e.g. [101],[137]) clearly help improve the performance of lip trackers. Typical shape constraints involve the characterisation of the possible lip shapes in terms of a linear subspace, such as the so-called ‘eigenlips’ [97] approach, that captures the main statistical modes of variation learnt from a ‘representative’ training set, by means of principal component analysis.

In this chapter, however, we show how an exclusively colour-based lip tracking system like the one described in [138] can be used as a bootstrap tracker. This tracking method is, thus, an intermediate system aimed at recovering typical lip shapes that would be used to learn the main modes of lip shape variation (chapter 5) so as to construct a more robust tracking system with adequate shape constraints. Both trackers are based on a B-spline representation of the lip contour, and on statistical chromaticity models to characterise respectively the lips and the surrounding skin area. B-splines have received considerable attention for tracking purposes [112, 13, 137] and are usually favoured with reference to other dynamic contour representations, such as snakes [52] because of their desirable properties like compactness, easy parameterisation and lower computational cost.

4.2 Statistical colour modelling

Cotton et al. [140] describe a model of colour formation within human skin based on the Kubelka-Munk theory of scattering and absorption within inhomogeneous materials\(^1\). By considering the skin to be a layered construction of such materials, the stratum corneum, epidermis, papillary dermis and reticular dermis, and by exploiting the physics related to the optical interface between these layers, the model generates all possible colours occurring within normal human skin. In particular, the model predicts\(^2\) that all skin colours have to lie on a simple curved surface patch within a three-dimensional color space bounded by two physiologically meaningful axes, one corresponding to the amount of melanin within the epidermis and the other to the amount of blood within the dermis.

This is of particular relevance for our purposes as the lips appear as a somewhat different histological structure from the surrounding skin resulting in their stereo-

\(^1\)see [141] for a review of the optical properties of biological tissues

\(^2\)as is reportedly verified by comparing the CIE LMS coordinates of a representative, cross-racial sample of fifty skin images with the LMS coordinates predicted by the model.
4.2. Statistical colour modelling

typical reddish colour tones. However, there are some noteworthy implications of the above mentioned work, supported by common experience, namely that, albeit the hue of skin colour lies in a close range of values, colour variations should be expected, not just across racial groups or personal identity, but also for a certain person over time, as a result of tanning, emotional reactions or any other induced temporary change of the layered structure (e.g. varying blood flow under the lips epidermis during speech production). For instance, as described in [142], the melanin (the pigment responsible for tanning) in the epidermis not only reduces the amount of back-scattered light, but also has a characteristic absorption spectrum that varies between individuals.

Whereas the skin surrounding the mouth appears as a somewhat planar uniform surface, lip shape is particularly variable. Because of this shape variability, lip intensity will also exhibit a similarly varying behaviour according to the direction of the illuminating source and the lip surface orientation. Colour models based on chromaticity [47] (i.e. intensity-normalised RGB components, as opposed to using directly RGB information) offer a certain resilience to illumination changes but, as typical cameras are usually non-linear, changes in illumination can also involve changes in colour appearance [143]. In addition to that, we can certainly expect to find specularities, saturation phenomena and coillumination effects which will contribute to further increase of colour variance.

4.2.1 Extraction of colour models

The original 3-dimensional RGB colour space is transformed into the 2-dimensional chromaticity space by normalising the red and green components of each pixel by its intensity ($L_1$ norm):

$$\begin{align*}
(r, g) &= \left( \frac{R}{\frac{1}{3}(R + G + B)}, \frac{G}{\frac{1}{3}(R + G + B)} \right)
\end{align*}$$

(4.1)

As mentioned in [144], the above colour normalisation achieves independence of the scene geometry (in the absence of highlights). On the negative side, the normalisation presents a non-removable singularity at the zero-signal point ($R=G=B=0$) and is highly unstable near this point. Furthermore, sensor signal-to-noise ratio is poor near that point, so that pixels near zero signal cannot be classified reliably. Likewise, it is worth mentioning that the chromaticity of two points on a surface of the same material will not be necessarily the same if one point is directly illuminated and the other is in a shadow, unless the light incident in the shadow has the same spectral characteristics as the direct illumination.

To account for individual variability and differences in illumination, the extraction of statistical colour models is done by supervised chromaticity clustering operated on a cropped rectangular window of the first image of the sequence to be tracked. The actual implementation of the clustering routine is based on the ISODATA algorithm described in [145]: a chromaticity histogram is computed and as many peaks as the number of clusters selected are determined as the starting cluster centroids. In an
iterative process that is run up to convergence the pixels are reclassified according to their distance to the current centroids, and these are then recalculated.

I should like to stress that, at this stage, the emphasis is not on getting a fine segmentation result, but in recovering the clustering trend of the colour information. Clustering algorithms that exploit both spatial coherence and feature space information (e.g. [146]), as opposed to only considering the feature space, are reportedly better for image segmentation. Although this might sound somehow contradictory, it is worth reminding that because of aiming at spatial coherence, the segmentation can include pixels that are not very representative of the region they have been classified into\(^3\), and, thus, are not that convenient to be used in colour model building.

In Figure 4.1 we can see how the above described process works. The mid figure shows the chromaticity histogram (brighter values represent higher frequency of occurrence) computed for the region of interest indicated in the left image with a rectangular window. After applying the clustering routine we can identify a first cluster corresponding to the skin area, a second one for the lips and a third one acting as a reject class. The positions of the centroids associated with the ‘skin’ and ‘lips’ clusters are marked in the represented chromaticity histograms with a black square and a white one respectively. Apart from the centroids, the clustering routine also outputs the covariance matrix associated with each cluster. These parameters (cluster centroid and covariance matrix) are then used to model each class (‘skin’ and ‘lips’) as unimodal bi-variate normal distributions:

\[
p(x|\omega_i) = \frac{1}{\sqrt{(2\pi)^2|\Sigma_i|}} \exp \left[ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right], \quad i = \{\text{lips, skin}\} \quad (4.2)
\]

\(^3\)This point was tested with a graph-theoretical clustering algorithm like that described in [146], which resulted in some clear artifacts because of enforcing spatial coherence when colour separability between the classes was poor.
4.3. Lip representation

Whereas the 'skin' class is reasonably well modelled by a single Gaussian cluster, the colour characteristics of the lips are not that well defined and, not surprisingly, several researchers have adopted a slightly more complex model for the latter, by representing the lips as a mixture of normal distributions [116, 139].

4.3 Lip representation

Lip contours are represented by closed (periodic) quadratic B-splines [121, 123] consisting of \( N \) spans and \( N \) control points \( q_i \) (\( i = 0, \ldots, N - 1 \)), \( q_i = (q_i^x, q_i^y)^T \).

B-splines allow for a compact representation of a contour in terms of a reduced number of control points. They provide a shape model with a certain shape selectivity that traditional snakes lack and their computational cost is also lower.

The position of a point \( r_i = (x_i, y_i)^T \) in the \( i \)th span is given by:

\[
\begin{align*}
  r_i(s)^T &= s^T M_i Q_i \\
\end{align*}
\]

where:

- \( Q_i = (q_{i-2}, q_{i-1}, q_i)^T \)
- \( M_i \) is a standard shape matrix [121]
- \( s = (1, s, s^2)^T \) is the parameter vector, \( 0 \leq s < 1 \).

In Figure 4.2 we can see a typical lip shape using a quadratic B-spline representation with 11 control points (black crosses in the image). The corresponding 11 spans are depicted in different alternate colours.

If we define a new control point vector space by stacking vector coordinates columnwise, first the \( x \) coordinates, and then the \( y \) coordinates, \( i.e. \)

\[
Q = (q_0^x, \ldots, q_{N-1}^x, q_0^y, \ldots, q_{N-1}^y)^T
\]

and consider that the parameter \( s \) now takes values in the range \( 0 \leq s \leq L \) (the length of the interval, \( L \), is equal to the number of control points \( N \) for closed splines), we can express the position of a point on the spline in a more compact way as:

\[
\begin{align*}
  \bf{r}(s) = U(s)Q &= \left( \begin{array}{cc}
  \bf{B}(s)^T & 0 \\
  0 & \bf{B}(s)^T \\
\end{array} \right) Q \\
\end{align*}
\]

where \( \bf{B}(s) \) is a vector of B-spline basis functions [113], where

\[
\bf{B}(s) = (B_0(s), B_1(s), \ldots, B_{N-1}(s))^T
\]
Chapter 4. The bootstrap tracker

Figure 4.2: Quadratic B-spline representation of a typical lip shape

Shape fit accuracy improves by increasing the number of spline spans. This will increase the computational cost and also the likelihood of tracking loss or unstable behaviour because of attempts made to fit unnecessarily fine detail, coupled to little support from the measurements effected for each span. Fit accuracy can also be improved by increasing the order of the spline. To this aim, cubic splines [147] are advocated in the literature. However, quadratic splines were considered sufficient because of exhibiting the same concavity/convexity properties in a single span. Consequently a lower bound on the appropriate number of spans or control points is given by the number of inflexions occurring along the contour. By visual inspection of typical mouth shapes during speech production, it was concluded that 10-11 control points would be enough to model the lip contour.

4.4 Spline initialisation

Chromaticity clustering or pixel classification can be applied in a cropped image containing the mouth region. This clustering process allows for the segmentation of the 'skin' and 'lips' areas. The segmented lip region is then used to generate an elliptic B-spline approximation to the lip contour. If pixel classification is preferred, the resulting spatial distribution can be used analogously to estimate the best fitting ellipse.

The ellipse parameters are estimated by the method of moments [32]. The center \((x_0, y_0)\) is computed as the centroid of the lips region:

\[
x_0 = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad y_0 = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

whilst the ellipse rotation angle \(\theta\) is computed as follows:
4.5. Lip boundary estimation

Eventually, the major and minor axes are found respectively by solving:

$$\tan 2\theta = \frac{2 \sum_{i=1}^{n} (x_i - x_0)(y_i - y_0)}{\sum_{i=1}^{n} (y_i - y_0)^2 - \sum_{i=1}^{n} (x_i - x_0)^2} \quad (4.8)$$

Eventually, the major and minor axes are found respectively by solving:

$$a = \left(\frac{4}{\pi}\right)^{\frac{1}{4}} \left[\frac{I_{\text{max}}^3}{I_{\text{min}}}\right]^{\frac{1}{8}}, \quad b = \left(\frac{4}{\pi}\right)^{\frac{1}{4}} \left[\frac{I_{\text{min}}^3}{I_{\text{max}}}\right]^{\frac{1}{8}} \quad (4.9)$$

where

$$I_{\text{max}} = \sum_{i=1}^{n} [(x_i - x_0) \cos \theta - (y_i - y_0) \sin \theta]^2, \quad (4.10)$$

$$I_{\text{min}} = \sum_{i=1}^{n} [(x_i - x_0) \sin \theta - (y_i - y_0) \cos \theta]^2 \quad (4.11)$$

Once the axes are computed an elliptic B-spline is generated, the control points of which are evenly spaced along the contour. Some examples of spline initialisation for some of the speakers can be seen in Figure 4.3.

4.5 Lip boundary estimation

Lip tracking proceeds by extracting profiles along the normals to a B-spline modelling the outer lip contour and estimating thereon the position of the lip boundary. Search along contour normals is a computationally low cost feature search procedure the suboptimality of which is usually justified on the grounds of the commonly referred aperture problem [131, 129, 148], which implies the impossibility of determining tangential displacements along the contour.

4.5.1 Estimation based on log-likelihood ratio

In the previous work described in [138] the decision criterion follows the Bayesian approach [149], selecting that class for which the \textit{a posteriori} probability is higher or, analogously, for which the likelihood ratio \( \lambda \) is greater than the unit:

$$\lambda = \frac{P(\text{lips}|\mathbf{x})}{P(\text{skin}|\mathbf{x})}, \quad (4.12)$$

so that the ‘lips’ class would be chosen if \( \lambda > 1 \) and ‘skin’ otherwise.

It is more practical to work in terms of the log-likelihood ratio \( \xi \):

$$\xi = \log \left(\frac{P(\text{lips}|\mathbf{x})}{P(\text{skin}|\mathbf{x})}\right) \quad (4.13)$$
Figure 4.3: Examples of B-spline initialisation in the M2VTS database
4.5. Lip boundary estimation

If we substitute for the above probabilities and further assume that the two classes are equiprobable, the decision criterion can be expressed as follows

\[ J(x) = J_{\text{lips}}(x) - J_{\text{skin}}(x), \quad J_i(x) = -\log |\Sigma_i| - M(x - \mu_i), \]  

(4.14)

\[ M(x - \mu_i) \] being the squared Mahalanobis distance of point \( x \) to the mean of class \( i \).

The decision rule would assign \( x \) to the 'lips' class if \( J > 0 \) and to the 'skin' class otherwise. Once we have the spline initialisation and the statistical models, lip boundary search can proceed by sampling chromaticity profiles (typically five per span) along the spline contour and computing the decision criterion \( J \). Figure 4.4(a) shows a chromaticity image with an elliptic spline superimposed. The search path at the mid point of the 3\(^{rd} \) span is represented in the same image. In Figure 4.4(b), we can see the values taken by the criterion function along the extracted profile (index numbering increases in the inward-outward direction). The outer lip boundary can be characterised by the last zero-crossing of the criterion function along the profile.

A lip tracker based on this method of estimating the lip boundary worked quite accurately where lip chromaticity characteristics were reasonably good. Otherwise, this simple estimation mechanism could fail and lead to inaccurate or even unstable tracking. It should be noted that the estimation criterion to characterise the position of the lip boundary based on the 'last' zero crossing of the loglikelihood ratio is quite sensitive to noise. Furthermore, the larger covariance matrix of the 'lips' class does not help locate precisely the boundary between both classes when the colour characteristics are too similar, the loglikelihood term being dominated by Mahalanobis distances with reference to the 'skin' class, characterised by lower colour variance.

4.5.2 Enhancing robustness of the lip boundary estimation

In this section a new estimation method is devised with the aim of overcoming the limitation of the previous estimator. Let us consider that we have 2 classes \( A \) and \( B \) corresponding to 'skin' and 'lips' colour respectively. A point will be considered to lie on the boundary between both regions if there exists a profile passing through such a point for which the pixels 'on the left' of that point \( \omega_{\text{left}} \) would belong to class \( A \) and those 'on the right' \( \omega_{\text{right}} \) would belong to class \( B \).

Assuming measurement independence, the probability that a point \( x_i \) belongs to the boundary between regions \( A \) and \( B \), which is represented as the event \( A|x_i|B \) would be given by:

\[ P(A|x_i|B) = P(w_{\text{left}} \in A, w_{\text{right}} \in B) = \prod_{\omega_{\text{left}}} P(x_j, A) \prod_{\omega_{\text{right}}} P(x_k, B) \]  

(4.15)

By neglecting \textit{a priori} region probabilities, and considering loglikelihoods instead:

\[ \lambda(A|x_i|B) = \sum_{\omega_{\text{left}}} \log p(x_j|A) + \sum_{\omega_{\text{right}}} \log p(x_k|B) \]  

(4.16)
(a) Chromaticity image ($\gamma = 6$) with overlaid B-spline and search path

(b) Criterion function value along the profile

Figure 4.4: Searching for the outer lip boundary
4.5. Lip boundary estimation

Edge detection will be posed as a maximum likelihood estimation problem along search profiles, which, for $L$-variate Gaussian class conditional probability distributions, is equivalent to minimising the sum of Mahalanobis distances between each pixel $x_j$ ($x_k$) and its assumed class mean vector $\mu_j$ ($\mu_k$), i.e.:

$$\xi(x_i) = \sum_{\omega_{\text{left}}} M(x_j, \mu_A) + \sum_{\omega_{\text{right}}} M(x_k, \mu_B)$$

(4.17)

The size of the sample in both the ‘left’ and the ‘right’ side of the candidate point is set to the same value $\Delta$ (typically set to four pixels, taking into account the average width of lips in the M2VTS database), which characterises the smoothness of the position estimator. Furthermore, for multivariate Gaussian distributions and independent measurements, $\xi$ follows a $\chi^2$ distribution with $L$ degrees of freedom, with $L = m(N_{\text{left}} + N_{\text{right}})$, where $m$ is the dimension of each $x_j$ ($x_k$) and $N_{\text{left}}$ ($N_{\text{right}}$) is the sample size on the left (right) of the candidate boundary point.

The role played by the sample size is the reduction of the estimator variance, thus increasing noise robustness, at the expense of higher computational cost.

Once again, we could be in the situation where the estimation is dominated by the Mahalanobis distances to the ‘skin’ prototype, i.e. it could happen that the boundary estimation is ‘shifted’, being supposed to lie in the skin area, because pixels of the ‘skin’ class might yield not too high Mahalanobis distances with reference to the ‘lips’ class, and pixels on the ‘skin’ side of the boundary estimate could be very close to their class prototype, resulting in very low Mahalanobis distances. As a result, the $\xi$ value corresponding to this situation might well be lower than that occurring at the real boundary.

This led us to consider the following heuristics to improve estimation robustness:

- Disregard the noisy, unreliable covariance matrix associated with the ‘lips’ class, and use that of the ‘skin’ class instead. Because of this manipulation, the Mahalanobis distances of ‘lips’ pixels are now emphasised.

- Conditional limitation of the Mahalanobis distance. Some pixels can be quite noisy but still be more likely to correspond to their alleged class than to the other. In this cases, if the Mahalanobis distance of this pixel is (presumably) too high, but still lower than that to the other class, its Mahalanobis distance is truncated to a statistically significant value. In practice, this means that, assuming bi-variate Gaussian distributions, the Mahalanobis distances (that would then follow a $\chi^2$ distribution of 2 degrees of freedom) will be lower than 6.0 with a probability of 0.95.

- When computing the criterion function for a given pixel, pixels along the profile on both its ‘right’ and ‘left’ hand side are considered, but not the very pixel itself, in order to avoid the influence of mixing effects in the estimation: because of the characteristics of the image capturing process, the sensor input corresponding to a point lying on the boundary of two regions, is likely to integrate signals of both regions.
Accordingly, $\xi(x_i)$ is redefined as follows:

$$\xi(x_i) = \sum_{j=1-\Delta}^{i-1} f(M_{\text{skin}}(x_j, \mu_{\text{lips}})) + \sum_{k=i+1}^{i+\Delta} f(M_{\text{skin}}(x_k, \mu_{\text{skin}}))$$

where

$$f(M_A(x, \mu_A)) = \begin{cases} 
M_A(x, \mu_A) & \text{if } M_A(x, \mu_A) < 6.0 \\
6.0 & \text{if } \begin{cases} 
M_A(x, \mu_A) < M_A(x, \mu_B) \\
\text{and} \\
M_A(x, \mu_A) \geq 6.0
\end{cases}
\end{cases}$$

\subsection*{4.6 Tracking mechanism}

As mentioned earlier, tracking proceeds by sampling profiles perpendicularly to the contour of the spline, and estimating new lip boundary positions. These measurements are then taken into account to estimate a new B-spline that will be the reference contour for the following frame.

Although no proper shape constraints are applied, a simple and unconstrained shape control mechanism can be provided by checking the projection of lip boundary points onto the main lip axis. Due to wrong measurements, the B-spline can adopt irregular shapes that would eventually result in tangling contours. Although the B-spline could be able subsequently to resume stable lip tracking, non-recoverable contour looping may occur. To avoid these situations (see Figure 4.5(a)), after each spline update, the projection onto the main lip axis of the vector joining each point along the lip boundary with one of the vertices is measured (Figure 4.5(b)). The projection should monotonically increase and have its maximum when the other vertex is reached. Spans that violate this condition are linearised (i.e. the origin and final points of the spans involved are substituted by straight lines) and the full B-spline is then reparameterised. Vertices are supposed to be situated at the two most horizontal locations of the lip boundary and are the basis for reparameterisation. The spline control points will be repositioned along both paths joining the vertices at equally spaced intervals. Each path can have a different number of control points to better represent shape variability for the lower and for the upper lip boundary. The final situation is represented in Figure 4.5(c).

\subsection*{4.6.1 Estimation with spatial and temporal smoothing}

Lip boundary search is done as described above. However, the final estimates of the actual position benefit from further robustness by considering the 2 best candidate positions for the boundary. According to tracking experience, the measurement probability distribution is not Gaussian. Because of poor colour contrast in some
areas, and because of the particularities of the lip geometry it would be possible to misestimate the position of the boundary on the sole grounds of the probabilistic reasoning already sketched. Paradoxically some of those ‘misestimates’ happen to occur on the actual lip boundary or close to it but in inappropriate locations. A good solution could consist of tracking the measured probability distribution as reported in [133]. This solution comes, however, at the expense of creating a potentially infinite search tree where all ‘probable’ minima are tracked. In the approach described herein the estimate is computed in a two-stage process: minima filtering and estimation update.

As for tracking dynamics, no artificial model has been assumed and new position updates are made after considering the probabilities of both the current estimate and the actual measurement, providing a natural way of smoothing tracking behaviour.

- Minima filtering: a number of estimator minima are retained and the estimation is formulated as the weighted average of their locations. The weights are measurements of likelihood and include both the measurement goodness (with reference to the colour model) as well as an estimate of its plausibility in terms of the distance from the current location. Assuming that these two factors are independent, the estimator can be formulated as follows:

$$\zeta = \frac{\sum P_{\text{colour}}(x_i)P_{\text{distance}}(x_i)x_i}{\sum P_{\text{colour}}(x_i)P_{\text{distance}}(x_i)}$$

(4.20)

where

$$P_{\text{colour}}(x_i) = 1 - P(\xi \leq \xi(x_i))$$

(4.21)

i.e. the probability\(^4\) that the colour-models related Mahalanobis distance (which follows a \(\chi^2\) distribution of \(2m\Delta\) degrees of freedom, \(m\) being the dimensionality of the selected colour space, i.e. 2 for chromaticity) is bigger than the current measure, and

\(^4\)It is possible to establish the parallelism between this probability term and the so-called single-comparison false-nonmatch rate defined in [150], which relates to the case where a sample is incorrectly not matched to a truly matching prototype because the distance between the two is greater than a fixed threshold.
\[ P_{\text{distance}}(x_i) = \frac{1}{\sqrt{2\pi\Delta}} \exp\left(-\frac{x_i^2}{2\Delta^2}\right) \]  

which is assumed to be a zero-mean Gaussian distribution of standard deviation equal to the size of the estimation window \( \Delta \). As for the number of minima, this is limited in practice to 2.

- **Final update:** The final estimation \( \hat{\zeta}_n \) relies on the current measurement \( \zeta_n \) as well as on the previous estimate \( \zeta_{n-1} \). The procedure consists once again of a linear combination of both estimates:

\[ \hat{\zeta}_n = \frac{P_{\text{colour}}(\hat{\zeta}_{n-1})\hat{\zeta}_{n-1} + P_{\text{colour}}(\zeta_n)\zeta_n}{P_{\text{colour}}(\hat{\zeta}_{n-1}) + P_{\text{colour}}(\zeta_n)} \]  

4.6.2 An alternative approach based on weighted least squares

Figure 4.6 shows the estimated lip boundary displacements along the current B-spline contour, i.e. profiles are sampled along the normal to the current spline contour at a number of points, and the expected position of the lip boundary is determined for each profile. The probability \( p_i \) of the displacement \( \Delta_i \) associated with the \( i \)-th sampling point along the contour is estimated as \( p_i = P_{\text{colour}}(\Delta_i)P_{\text{distance}}(\Delta_i) \), with \( P_{\text{colour}} \) and \( P_{\text{distance}} \) as defined above.

As can be seen (taking into account the starting point shown by the top left image in Figure 4.8), the measured displacements seem to take reasonable values except in some significant cases, where the magnitude of the estimated displacements appears extraordinarily high. This can be better appreciated if we compute the displacement histogram, as shown in Figure 4.7.

The coordinates of the sample points extracted along the contour can be expressed as a linear combination of the coordinates of the B-spline control points. For instance, for the \( x \) coordinates, we would have

\[ x = A q_x \]  

where \( x \) is of dimensionality \( n = N \times d \), \( N \) being the number of control points and \( d \) the number of sampling points for each span of the spline.

On the other hand, if we have obtained a number of \( n \) boundary estimates by uniformly sampling the B-spline contour, we could estimate the control point coordinates as:
4.6. Tracking mechanism

Figure 4.6: Estimated lip boundary displacement with associated probability estimates (×20)

Figure 4.7: Histogram of estimated lip boundary displacements
Chapter 4. The bootstrap tracker

\[ \hat{q}_x = (A^T A)^{-1} A^T x \]  

(4.25)

This would be equivalent to the solution of a least squares problem where all measurements are valid. Invalid measurements could be excluded from the estimation and \( \hat{q}_x \) would be estimated as:

\[ \hat{q}_x = (A_r^T A_r)^{-1} A_r^T x_r \]  

(4.26)

where \( x_r \) represents a vector of dimensionality \( r, r < n \), as wrong measurements are suppressed, and, likewise, \( A_r \) is the matrix obtained through suppression of the rows corresponding to the rejected measurements.

In a more general scenario, measurements could be weighted according to their goodness. If we consider a diagonal weighting matrix \( W \), the estimation of the control point coordinates would be given by:

\[ \hat{q}_x = (A^T W^T W A)^{-1} A^T W^T x \]  

(4.27)

It is easy to show that Equation 4.26 is a particular case of Equation 4.27 where the weights are 1 for 'good' measurements and 0 for 'bad' measurements.

The new estimation methods just described are compared with the old one in Figure 4.8. The old method (top right image) is similar to Equation 4.25, but where the measurements had been 'smoothed' previously by generating the estimates as a weighted linear combination of the current location and a weighted average of the best two estimates along the sampled profile. In the bottom left image, we can see what happens when we directly compute the pseudoinverse without taking into account measurement plausibility. The net effect is that a very reduced number of bad measurements affects the whole estimation and the resulting new B-spline contour is clearly distorted. If, on the other hand, measurements are weighted with their corresponding probability estimates, the results shown in the bottom right image are obtained. In principle, no significant differences can be established between these latest results and those obtained with the old estimation method. However, the old method showed a certain tendency to drift away from the lip contour if 'distracted' long enough, as whether or not the measurements are good for a given profile, it is still committed to providing a weighted estimate of the most likely contour locations. Its focus is, thus, local and does not put measurements in a general context, as Equation 4.27 does, where contributions to the control point estimates are commensurate to their plausibility according to the derived statistical models.

In Figure 4.9 we can see the results of some tests consisting of the introduction of artificial, faulty measurement and measurement probability modification to suppress wrong measurements in the estimation process. The top left figure shows the effect of a single poor measurement in the control point estimation (Equation 4.27). An artificial displacement of +15 pixels (the image size is 65 x 35 pixels) was applied to the \( Y \) coordinate of one of the measurements, and its probability artificially set
4.7 Tracking results

The bootstrap tracker was tested on the M2VTS database [57], on sequences consisting of frontal images of 37 different speakers uttering the sequence 0 to 9 in French. Each sequence corresponds typically to about 100-200 images (between 4 and 8 seconds of speech) of 350 x 286 pixels.

Tracking was stable for all the subjects, although some contour inaccuracies are apparent in some frames. Because of only exploiting colour information, lip bound-
Figure 4.9: Testing the effect of outliers and the suppression of poor measurements.

ary estimation can be misled when the colour characteristics of each region are no longer that different. As an example, we can see in Figure 4.10 the tracking results obtained for every tenth frame of the sequence ‘sp.04.v’. The results can be better appreciated in the following Figure 4.11, where all the frames of the sequence are represented. As can be seen, contour fit accuracy is quite good in general. However, at some points the spline gets a bit wiggly and goes off the expected contour, typically into the lip region. This can be noticed, for instance, in a few of the images of the bottom row of Figure 4.11. The colour measurements close to the outer boundary of the lower lip do not correspond that well to the lips statistical model, and are rather similar to the ‘skin’ class. As a result, the lip boundary estimation is shifted towards the lip area, as there is no shape to adapt to\(^5\). Some other problematic points are the corners of the lips, where the lips become thinner (less measurement support) and their colour components darker (thus taking chromaticity values that depart from the estimated colour model). Nonetheless, for these sequences corresponding to speech utterance, the tracker managed to latch onto the ‘right’ contour once the colour measurements turned back to ‘normality’.

4.8 Discussion

We have presented a colour-based tracker as an intermediate step towards building a more robust one by making use of appropriate shape constraints. The conducting line of this development is a bottom up approach where the amount of \textit{a priori}

\(^5\)Coupled templates are well known in the literature [129, 13] as a way of reinforcing tracking stability.
Figure 4.10: sp.04.v: tracking results for every tenth frame, in raster order
Figure 4.11: sp_04.v: tracking results for every frame (0-112), in raster order
4.8. Discussion

constraints or manual intervention is kept to a minimum, to eventually come up with an adequate set of modes of lip shape variation.

The fact that we obtain a majority of 'valid' shapes with the described bootstrap tracker validates the proposed approach. Note that (other than the tracking inaccuracies described above) the rest of the contour inaccuracies can be ascribed to high frequency noise (reflecting the lack of adequate shape smoothing) that will statistically average out when the whole ensemble of shapes are considered. As described in the following chapter, representative modes of lip shape variation can be automatically extracted from the above tracking results.
Chapter 4. The bootstrap tracker
Chapter 5

Robust estimation of main modes of lip shape variation

5.1 Introduction

The estimation of main modes of shape variation is a typical chicken and egg problem and not surprisingly it is common to resort to some sort of manual intervention such as the selection of those ‘representative’ images or even the manual registration of reference points (e.g. [101]). Although relevant works exist in analogous domains where the shape extraction can be done automatically [131], the common assumption was the availability of a ‘clean’ training set without outliers that would otherwise distort the classical estimation of the shape parameters.

Along these lines, it is shown here how an exclusively colour-based lip tracker can be used as a bootstrap procedure to robustly retrieve the main modes of variation of lip shapes. Although colour information is in itself insufficient and the tracker may yield a considerable number of wrong lip shapes, it is shown here how, under the assumption that a majority of tracked shapes are valid, these so-called eigenlips can be computed through a robust estimation of the lip shape covariance matrix.

The approach followed represents a novel application of robust statistics to computer vision. Using the affine-normalised output of the bootstrap tracker, a new method is presented to robustly estimate the principal modes of variation of the ensemble of lip shapes in a fully automatic, unsupervised way, and obtain a low-dimensional representation of the space of acceptable lip contours.

5.2 Robust covariance estimation

In order to have an adequate eigenspace representation of the data, it is of central importance to have robust estimates of both the mean (location) and the covariance matrix (scale) of the data. A major global robustness measure is given by the breakdown point (see [151] for a formal definition and further robustness measures), which
can be loosely defined as the fraction of contaminated data required to make the estimator fail. For instance, the breakdown of the classical average-based location estimator is $1/n$, where $n$ is the number of data, as just one outlier measurement is enough to make the estimator significantly depart from what would be otherwise expected. Analogously, the breakdown point of the median estimator is 0.5, since this estimator would tolerate up to 50% of contaminated data. A breakdown of 0.5 is actually the theoretical maximum achievable since for higher values it would become impractical to recognise the ‘good’ data from amongst a sample set comprising a significant fraction of outliers with an unlikely but theoretically possible ‘anomalous’ distribution. In order to better illustrate this point, imagine a situation where more than half of the data are outliers of the model we try to estimate, and that the outliers happen to have a well defined cluster structure. Quite likely, our estimator would come up with a ‘robust’ estimate of the cluster of outliers instead of the one we are after.

It is worth stressing [152] that the robustness of an estimator can be "measured" by several properties besides its breakdown point, and that there are other properties that can be as desirable at the very least. In particular, two properties of great importance are that the performance of a method should deteriorate only slightly under small deviations, and that it should have a good estimation efficiency (accuracy). Two main reasons why the need for a very high breakdown is sometimes irrelevant are a) that the formal definition of breakdown point is rather extreme in that it requires the estimate to be arbitrarily far from the actual value, and b) that the breakdown point takes into account the worst possible scenario, probable but possibly absurd.

Independently of the above considerations, and turning back to our problem, it is assumed that we have a significant number of valid tracked lip shapes, but also a considerable (yet smaller than the number of ‘good’ data) amount of incorrect, and possibly severely distorted, tracking results, and the high breakdown point has prevailed as the main robustness criterion being sought after.

In [151], a robust multivariate estimator of both location and scale with a high breakdown point of 0.5 is suggested. The estimation is quite intuitive and implies finding the affine transformation $z = L^{-1}(x - \hat{\mu})$ (where $x$ is an input vectors, and $\hat{\mu}$ and $\hat{L}$ are respectively mean and covariance estimates), such that the following objective is achieved: the projected points have zero mean, the covariance matrix is $\beta I$ ($\beta$ is a coefficient slightly lower than unit value), and half of the points lie within the hypersphere of radius $\sqrt{m/\beta}$, where $m$ is the dimensionality of the subspace.

The algorithm can be posed as an optimisation problem in an $m(m + 3)/2$ dimensional space (the $m$ elements of the mean vector, plus the $m(m + 1)/2$ of the covariance matrix), which for $m = 10$ already involves 65 dimensions.

Another robust estimator being referred to in [151], also with a 0.5 breakdown point is MVE, which stands for minimum volume ellipsoid. The method consists of determining the ellipsoid with the smallest volume that contains (at least) 50% of the data, and uses its centre as a location estimate. An analogous method by the same author is MCD (for minimum covariance determinant), which exhibits
better estimation efficiency and for which a stochastic version with quite a low computational cost exists. The original algorithm [16] and its adaptation for its application to the robust estimation of eigenlips is described in the following sections.

5.3 Minimum Covariance Determinant Estimator

The objective of the MCD estimator is to find \( h \) observations (out of \( n \)) whose classical covariance matrix has the lowest determinant. The MCD estimate of location is then the average of those \( h \) points whereas the MCD estimate of scatter is their covariance matrix. The resulting breakdown value equals that of the MVE, but presents better statistical efficiency and a faster convergence rate. Robust distances based on the MCD are more precise than those based on the MVE, and hence better suited to expose multivariate outliers.

The MCD algorithm consists of the following steps:

1. Random sampling: extract a random \((m + 1)\)-subset \( J \), and then compute \( T_0 = \text{ave}(J) \) and \( S_0 = \text{cov}(J) \). If \( \det(S_0) = 0 \), extend \( J \) by adding another random observation, and continue adding observations until \( \det(S_0) > 0 \).

2. Compute the distances \( d_0(i) = (x_i - T_0)^T S_0^{-1} (x_i - T_0) \), for \( i = 1, \ldots, n \). Sort them into \( d_0(\pi_0(1)) \leq \ldots \leq d_0(\pi_0(n)) \) and create \( H_1 = \{ \pi_0(1), \ldots, \pi_0(h) \} \), where \( \pi_0(m) \) stands for the sample (amongst the whole data set) associated with the \( m \)-th lowest distance after the 0-th algorithm iteration.

3. Iterate until convergence\(^1\):

   - Compute \( T_k = \text{ave}(H_k) \) and \( S_k = \text{cov}(H_k) \)
   - Compute the distances \( d_k(i) = (x_i - T_k)^T S_k^{-1} (x_i - T_k) \), for \( i = 1, \ldots, n \). Sort them into \( d_k(\pi_k(1)) \leq \ldots \leq d_k(\pi_k(n)) \), where \( \pi_k(m) \) stands for the sample associated with the \( m \)-th lowest distance after the \( l \)-th algorithm iteration, and create \( H_{k+1} = \{ \pi_k(1), \ldots, \pi_k(h) \} \)

4. Selection: Set \( T = T_k \) and \( S = S_k \)

With reference to random sampling, the same considerations apply as in the RANSAC paradigm [153] regarding the probability \( p \) of having at least one ‘clean’ \((m + 1)\)-subset among \( r \) random \((m + 1)\)-subsets:

\[
p = 1 - (1 - (1 - c)^{m+1})^r
\]  \hspace{1cm} (5.1)

where \( c \) is the fraction of ‘contaminated’ pixels.

\(^1\)In [16], it is proved that the sequence \( \det(S_k) \) verifies \( \det(S_1) \geq \det(S_2) \geq \ldots \), and as it is non-negative it must converge. Since there is a finite number of \( h \)-subsets, there must be an index \( t \) such that \( \det(S_t) = 0 \), which corresponds to the case where the data perfectly lies in a subspace of lower dimensionality, or \( \det(S_t) = \det(S_{t+1}) \), which is the most typical scenario. Usually, \( t < 10 \) in practice.
Nevertheless, because of the favourable convergence properties of the algorithm, it is not necessary to sample the data so densely as in RANSAC, and not more than 500 \((m + 1)\)-subsets, with \(m\) of the order of 20, are typically sampled. The pair \((T, S)\) with the lowest covariance determinant amongst the selections made for each random subset is chosen as the solution of the estimation problem.

The default value of \(h\) is \([(n + m + 1)/2]\), for which the maximum breakdown value \((n - h + 1)/n\) is achieved. However, it is possible to select any other value such that \([(n + m + 1)/2] \leq h \leq n\) and, if the data is ‘sure’ to contain less than 25\% of contamination, a good compromise between breakdown value and statistical efficiency is obtained by setting \(h = 0.75n\).

### 5.4 MCD for robust eigenlip estimation

One of the purposes in computing eigenlips for shape representation is dimensionality reduction. In principle, if the original dimensionality of the input data is \(d\), we can readily implement the MCD estimator in the original \(d\)-dimensional space, and then retain the first \(m\) modes of variation corresponding to a required percentage of the total variability. However, the fact that the \(d\)-dimensional estimates are the best robust ones we can get does not necessarily mean that the first \(m\) eigenvectors are the best \(m\)-dimensional representation we could get.

If we assume that the data can be approximated by a multivariate (say \(t\)-variate) Gaussian distribution, the Mahalanobis distance between the samples and the mean vector will follow \([154]\) a chi-square distribution with \(t\) degrees of freedom. The situation is equivalent to considering that the data can be affine transformed into a zero mean hypersphere where each ‘principal axis’ would have unit variance. In Figure 5.1, we can see the effect of modifying the required data dimensionality or the fraction of required data.

If we assume that the ‘intrinsic’ data dimensionality is \(m\), we can now see, from Figure 5.1, the effect of ‘refining’ data representation by considering a higher dimensionality. The direct implication is a growth in the radius of the hypersphere we would need to retain the same fraction of the data required for the estimation. The ‘importance’ of each ‘principal axis’ is, however, the same. As a result, the price to pay for a lower representation error is a higher risk of distorting the covariance estimate through getting more outliers into the region used to calculate the parameters.

Hence, the preferred approach to compute the eigenlips has consisted of running an MCD estimator for each possible dimensionality and then keeping the one with the lowest dimensionality for which a certain percentage of the total variance is obtained (say 90-95\%). To accomplish this, we have modified the MCD estimator in the following way:

- First, in both steps 2 and 3, before the computation of the Mahalanobis distance, the eigenvectors and eigenvalues of \(S_k\) are computed and each input
5.4. MCD for robust eigenlip estimation

A low-dimensional eigenlip representation of the lip shapes is just an approximation associated with a certain reconstruction error because of the neglected components responsible for the residual data variance not spanned by the selected number of eigenlips. If the estimation were based only on in-space (i.e. in the linear subspace spanned by the selected number of eigenlips, or factor space, as commonly termed in the multivariate statistics literature) Mahalanobis distances, it is likely that the estimation would fail because of the presence of outlier data that behave well in the subspace but have significant components in the orthogonal subspace so that they do not correspond to the assumed model where the variation outside the eigenlip subspace is marginal. As done in [41] for maximum likelihood detection of face-like objects, it is possible to assume that all components outside the eigenlip subspace can be considered as an uncorrelated zero-mean noise term of variance $\lambda^*$ which is related to the residual eigenvalues of the covariance matrix. If the dimensionality of the orthogonal subspace is large enough, the probability density function associated with that noise term can be assumed normal, according to the central limit theorem.

Figure 5.1: Effect of data dimensionality and the required fraction of data on hypersphere size

- Vector is projected onto the eigenvector subspace: $\tilde{x}_i = T_k + \sum_{j=1}^{m} a_j^{(i)} e_j^{(k)}$ is the projection of $x_i$ in the eigenspace, with $e_j^{(k)}$ being the $j$-th eigenvector of $S_k$ and $a_j^{(i)}$ the associated coefficient.

- Also in steps 2 and 3, the Mahalanobis distances $d_k(i)$ are defined as: $d_k^{(i)} = \sum_{j=1}^{m} a_j^{(i)2} \lambda_j^{-1} + \frac{\lVert x_i - \tilde{x}_i \rVert^2}{\lambda^*}$ where $\lambda_j$ is the $j$-th eigenvalue of $S_k$, and $\lambda^*$ is a variance term to normalise data approximation errors (in practice, $\lambda^* = \lambda_m$)
Chapter 5. Robust estimation of main modes of lip shape variation

[154]. This is not exactly the case and, furthermore, the residual eigenvalues can be quite dissimilar in magnitude, so that it has been finally preferred to overestimate the 'noise' variance by assuming that it is similar to the lowest eigenvalue \( \lambda_m \) of the eigenlip subspace.

A final estimation refinement for the sake of consistency when the data come from a multivariate normal distribution is the scaling of the covariance matrix estimate as follows [16]:

\[
S_{MCD} = \frac{mcd_i d'TS(i)}{\chi^2_{d+1,0.5}} S
\]

(5.2)

where the numerator refers to the median distance obtained with the estimated parameters, and \( \chi^2_{d,x} \) is the quantile of a chi-square distribution of \( d \) degrees of freedom for which the probability is lower than or equal to \( x \).

In [16] a further refinement would consist of performing a one-step reweighted estimate. This additional step boils down to recomputing location and scale estimates in the classical way using the data that verify \( d_{TS_{MCD}}(i) \leq \sqrt{\chi^2_{d+1,0.975}} \). Nevertheless the results of tests performing reweighted estimates did not actually show improvements and this approach is not used in the final estimation algorithm.

5.5 Experimental results

The tracking results of the 4th shot of the M2VTS database, totalling \( n = 5450 \) lip shape samples of 36 different subjects, were used to estimate the covariance matrix. Each sample comprises the coordinates of the 11 control points describing each B-spline curve, thus resulting in a 22-dimensional space.

The MCD estimator was run for all dimensionalities \( d = 2, \ldots, 19 \), and for two different fractions of data \( h \): \( h = 0.5n, h = 0.75n \). As an example, the resulting distance histograms together with the theoretical histograms (chi-square distributions) are shown in Figure 5.2. Recalling the considerations made in the preceding section, we can see how the goodness of fit (should the data be adequately modelled by a multivariate normal distribution) actually degrades with increasing values of the dimensionality.

As we increase the dimensionality of the approximating linear subspace, we are able to account for more and more of the total data variance. For instance, with \( d = 5 \), for which we get a very good agreement with the hypothesised chi-square distribution (i.e. the theoretical distribution for normally distributed clusters), the spanned variation is roughly 70% of the total variance, whereas with a 12-dimensional representation we can account for more than 95% of the total variance. However, we can also see how the distance histograms also depart more and more from the theoretical distributions with increasing values of the selected dimensionality for the approximating subspace. One of the main reasons for this can be found (see section 5.4) in the bigger radius of the \( d \)-dimensional hypersphere which is necessary to retain the
same fraction of data as the dimensionality increases. As a result, the probability of including outliers in the computation of the covariance matrix increases with the dimensionality of the representational subspace.

As an example of the performance of the algorithm we can see \textit{a posteriori} some of the lip shapes that are now rejected when estimating the covariance matrix. We can see how these shapes that exhibited abnormally large Mahalanobis distances are in fact hardly recognisable as acceptable. The bottom row of Figure 5.3 actually shows a few examples of these shapes, that are rejected by the algorithm. All of them have Mahalanobis distances above 20.0. On the other hand, some typical 'valid' shapes (in this case with Mahalanobis distances of about 1.0) are shown in the top row of the same figure.

### 5.6 Eigenlips computation

The variability we have referred to so far is that of the spline control points themselves, which does not necessarily correspond \cite{113} to the actual shape variability. In fact, although the lip outline is completely determined if we know the coordinates of the control points governing the spline, the control points themselves can be too weak a representation for verification purposes in the light of some of the limitations pointed out in \cite{132}, \textit{i.e.}:

- Apparently similar shapes may have quite different nodal representations due to variations in the placement of the control points.
- Arc-length parameterisation does not ensure that physically corresponding points will always have the same parameter values.

These problems are alleviated, on one hand, by enforcing consistence in control point numbering, and, on the other hand, with the current uniform parameterisation. There stills subsists, though, the inappropriateness of using control point distance as the metric considered for measuring actual shape variability. However, the main modes of variation associated with 'real' shape variation can be, anyway, computed from the robust estimate of the covariance matrix multiplied by a metric matrix \(H\).

Indeed, if we recall that a point \(\mathbf{r}(s)\) of the B-spline is defined by

\[
\mathbf{r}(s) = \mathbf{U}(s)\mathbf{Q} = \begin{pmatrix} \mathbf{B}(s)^T & 0 \\ 0 & \mathbf{B}(s)^T \end{pmatrix} \mathbf{Q}
\]

where \(s\) is the B-spline parameter \((0 \leq s \leq L)\), \(\mathbf{Q}\) is a control point vector, and \(\mathbf{B}(s)\) is a vector of B-spline basis functions, and further if we define an \(L_2\) norm \cite{113} for B-splines curves which is induced by the Euclidean distance in the image plane as

\[
||\mathbf{Q}||^2 = \mathbf{Q}^T \mathbf{H} \mathbf{Q}
\]
Figure 5.2: Mahalanobis distance histograms and theoretical chi-square distributions
5.6. Eigenlips computation

Figure 5.3: Top row: typical valid shapes; bottom row: shapes rejected by the algorithm

with

$$H = \frac{1}{L} \int_0^L U(s)^T U(s) ds = \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix}$$

the eigenlips can be computed [113] as the eigenvectors of $S_{MCD}H$.

For the B-spline model (periodic, quadratic spline of 11 control points) used to represent the contour of the lips $C$ is an $11 \times 11$ sparse circulant matrix:

$$C = \begin{pmatrix}
0.55 & 0.217 & 0.008 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.008 & 0.217 \\
0.217 & 0.55 & 0.217 & 0.008 & 0 & 0 & 0 & 0 & 0 & 0 & 0.008 \\
0.008 & 0.217 & 0.55 & 0.217 & 0.008 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.008 & 0.217 & 0.55 & 0.217 & 0.008 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.008 & 0.217 & 0.55 & 0.217 & 0.008 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.008 & 0.217 & 0.55 & 0.217 & 0.008 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.008 & 0.217 & 0.55 & 0.217 & 0.008 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.008 & 0.217 & 0.55 & 0.217 & 0.008 \\
0.008 & 0 & 0 & 0 & 0 & 0 & 0.008 & 0.217 & 0.55 & 0.217 \\
0.217 & 0.008 & 0 & 0 & 0 & 0 & 0 & 0.008 & 0.217 & 0.55 \\
\end{pmatrix}$$

The sparsity reflects the fact that each control point only affects the generation of the spline over a limited interval of the parameter $s$. 
In Figure 5.4 we can see the effects associated with a variation of ±2 standard deviations over the mean lips shape (represented in the centre of the picture) for each of the 5 principal axes of the selected 5-dimensional subspace. It is interesting to see that now, with the newly defined $L_2$ norm used to measure lip shape variation, the new subspace derived from the previously estimated 5-dimensional control point covariance matrix amounts in this case to about 85 % of shape variability.

5.7 Summary

A robust method for the computation of modes of shape variation from contaminated data provided by a colour-based lip tracker has been presented. The method achieves a low dimensional characterisation of lip shapes in terms of a reduced number of eigenlips by means of an adapted implementation of Rousseeuw's MCD algorithm [16]. By simply assuming a simple majority of valid lip shapes in the training set, meaningful modes of lip shape variation are obtained. The comparison of the resulting Mahalanobis distance distribution with the theoretical chi-square curves also shows the validity of a unimodal cluster model for the data, the more so the lower the dimensionality of the eigenlip space, where the lip shape cluster can be closely modelled by a multivariate normal distribution. Eventually, a 5-dimensional subspace accounting for about 85 % of the total shape variability can be selected as the robust solution to our spatial constraint estimation problem and be subsequently used in a robust tracking systems where only valid (i.e. lying in the shape space spanned by the computed eigenlips) lip contours are allowed.
5.7. Summary

(a) 1+  (b) 2+  (c) 3+  (d) 4+  (e) 5+

(f) mean lips shape

(g) 1-  (h) 2-  (i) 3-  (j) 4-  (k) 5-

Figure 5.4: Robust 5-dimensional eigenlips space
Chapter 5. Robust estimation of main modes of lip shape variation
Chapter 6
Shape-constrained lip tracking

6.1 Introduction

The approach followed for tracking that is described in this chapter can be termed shape-constrained tracking by synthesis. Based on the assumed shape and dynamics models, and on the prior distributions, a prediction for the following frame can be formulated and tested against the particular image. In our case, we look for affine-transformed shape parameters that result in the lowest probabilistically-weighted reconstruction error given the extracted observations for the current state. Although, strictly speaking, the breakdown point [151] of this estimation procedure based on all observations available is very low (theoretically a single but substantially corrupted measurement is enough to significantly distort the estimation), this possibility is only of theoretical interest since the probabilistic weighting scheme relates the contribution of each observation towards the estimate to its probability and, therefore, the influence of gross measurement errors can be considered negligible.

A simple first order model is used for the dynamics. Accordingly, the estimation for the current frame is used as a prediction for the following one. The results obtained in the tracking experiments conducted on the M2VTS database [57] show significant improvements with reference to the exclusively colour-guided bootstrap tracker of chapter 4 and indicate that this simple model is enough to characterise lip motion during speech production. Nevertheless, more complex dynamics models are also discussed, as well as other possible improvements and refinements.

6.2 Shape models

The model chosen for characterising the lip contour consists of all possible affine-transformed B-splines whose control points are a linear combination of the eigenlips obtained using the robust estimation method that was presented in chapter 5.

Thus, we consider first an $m$-dimensional subspace generated by the eigenlips $e_k$ ($1 \leq k \leq m$, with $m = 5$ in practice), so that the coordinates of the affine-normalised B-spline control point vector ($N$ spans) are given by
where $Q^0_x$, $Q^0_y$ stand, respectively, for the $x$ and $y$ coordinates of the $N$ control points, $(Q^T_x Q^T_y)^T$ is the mean control point vector, and $b_1, \ldots, b_m$ are the coefficients of the expansion in eigenlips.

If we now consider an affine transformation matrix $A$,

$$A = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix}$$

(6.2)

the $Q_x$ and $Q_y$ coordinates of the affine-transformed B-spline control points can be computed as follows, making using of homogenous coordinates:

$$\begin{pmatrix} Q_x \\ Q_y \\ 1 \end{pmatrix} = \begin{pmatrix} Q^0_x \\ Q^0_y \\ 1 \end{pmatrix} A^T$$

(6.3)

The approach implicitly assumes a unimodal multivariate normal distribution for the coefficients of the expansion in eigenlips and a uniform distribution for the parameters of the affine transformation.

It is worth stressing the differences with regard to other similar models used in lip tracking. In [101], the eigenlips are computed without a previous affine-normalisation of the input data. This has the advantage of generating a more compact shape representation model, since all affine variability is now absorbed by the subspace and the coefficients of the expansion in eigenlips now span the whole shape variability existing in the training database. There are, however, some drawbacks. It is only when all lip images are perfectly aligned that affine variability is also characteristic of proper shape changes. If this does not hold, affine variability is likely to correspond to changes in image conditions such as different distance to the camera, facial rotation, perspective effects which do not necessarily correlate to ‘real’ changes in lip shape. Furthermore, it has been found that affine changes can account for most of shape variability. When absorbing affine variability in the construction of the eigenlip subspace, this ‘major’ source of variability is masking more subtle detail which is not reflected by simply elongating, rotating, etc a reference shape.

On the other hand, in [113] a linear subspace is considered that models affine deformations of a reference template. Blake et al (e.g. [139]) have used lip tracking for bimodal speech recognition but not for speaker recognition/verification. As a result, shape specificity is not a major concern. Although using more reference templates (the so-called ‘key-frames’) is a possibility whilst still keeping a strictly linear representation, the different affine deformations of each key-frame must be accounted for, resulting in vector correlation and, eventually, in a rather wasteful representation model.
6.3 Dynamics model and tracking mechanism

Following the preliminary results obtained with the bootstrap tracker (see chapter 4, a simple, first order model was considered enough for characterising lip dynamics during speech production. This can be represented as follows, using state space notation [124]:

$$\theta_n = \theta_{n-1} + v_n$$  \hspace{1cm} (6.4)

where $\theta_n = (b_1, \ldots, b_m, a_1, \ldots, a_6)^T$, $a_i$ ($1 \leq i \leq 6$) being each one of the 6 parameters of the affine transform operated on the B-spline generated by the linear combination of eigenlips given by the $b_j$ ($1 \leq i \leq m$) coefficients, and $v_n$ is a noise term.

With this simple model, the current state is used as a prediction for the following frame, i.e. $\hat{\theta}_n = \theta_{n-1}$. At that point measurements are extracted along the B-spline contour in order to generate an a posteriori estimate. The process of extracting measurements consists of sampling profiles perpendicular to the contour of the B-spline prediction in the same way as for the bootstrap tracker. In this way we obtain a set of measurements $\hat{z}_1 \ldots \hat{z}_{dN}$ ($d$ represents the sampling density, i.e. the number of profiles cast per span), with corresponding probabilities $p_1 \ldots p_{dN}$. The length of the sampling profiles is dynamically adjusted taking into account the median absolute deviation (MAD) of the lip boundary estimates with reference to the current position on the B-spline contour, which is analogous to the spatial gating mechanism described in [112].

Parameter estimation for the current frame is posed as the problem of calculating $\hat{\theta}_n^* = (b_1^*, \ldots, b_m^*, a_1^*, \ldots, a_6^*)$ that best fits the extracted measurements.

In principle, we can consider an error function $J_1(\theta)$ using a probabilistically-weighted Euclidean metrics:

$$J_1(\theta) = \sum_{i=1}^{dN} p_i ||z_i(\theta) - \hat{z}_i||^2$$  \hspace{1cm} (6.5)

However, in order to account for the prior shape distribution as well as not to become dependent on a particular sampling density or subspace dimensionality, the following criterion function $J$ was chosen:

$$J(\theta) = \frac{\sum_{i=1}^{dN} p_i ||z_i(\theta) - \hat{z}_i||^2}{\sum_{i=1}^{dN} p_i} + \frac{1}{m} \sum_{j=1}^{m} \frac{b_j^2}{\lambda_j}$$  \hspace{1cm} (6.6)

where $\lambda_j$ stands for the eigenvalue associated with the $j$-th eigenlip, and the second term of $J$ is nothing else but the dimensionality-normalised Mahalanobis distance of the affine-normalised lip shape to the mean shape. Therefore, the a posteriori estimate for the current frame is given by the $\theta_n^*$ that minimises $J$. 
It can be argued that the criterion function used to perform lip parameter estimation is not robust in a strict sense. Indeed, its breakdown point \(1/dN\). However, the probabilistic weighting (see chapter 4) penalises outliers, so that their influence is kept to a minimum, and preserves, on the other hand, the statistical convergence of least squares. Under normal operational conditions where the high breakdown point criterion can be considered too extreme this criterion function can still be considered 'robust' in Huber sense \([152]\): "the performance of a method should deteriorate only slightly under small deviations, and it should have a good accuracy".

Certainly, as done in the robust regression problem faced when dealing with the eigenface representation of occluded facial images (see Appendix C), and provided we also have an appropriate error model, we could devise an analogous solution to the problem of estimating the affine and shape parameters that best represent the lip boundary in each frame. Once again, it is not possible to establish a priori the number of 'corrupted' measurements, which is, furthermore, a figure that could change from frame to frame during tracking. We could obviously sit back and adopt an LMedS (least median squares \([155]\)) algorithm, but when the number of valid measurements is clearly higher than 50%, this estimator, however robust, shows poor statistical convergence as seen in the aforementioned face occlusion problem: the 'best' half of the measurements would be overfitted whereas otherwise valid measurements would be disregarded, thus resulting in poor overall accuracy.

As the a posteriori estimate is supposedly close to the prediction minimising \(J\), its estimation can be posed as a local optimisation problem, for which a Dynamic Hill Climbing \([156, 157]\) algorithm appears quite suitable, also in the light of the results obtained when solving robust regression problems of high dimensionality (see Appendix C). Dynamic Hill Climbing incorporates some heuristics that are broadly applicable in gradient-ascent problems, namely:

- the size of probing steps is adjusted to suit the local of the terrain, shrinking when probes do poorly and growing when probes do well;
- keeping track of directions of recent success, so as to probe preferably in the direction of most rapid ascent (descent);
- remembering the location of local maxima and restarting the optimisation routine at a place distant from previously located local maxima.

The combination of the above three heuristics results in an efficient optimisation routine with improved resilience to getting stuck in local maxima (minima) or terrain ridges.

\(^{1}\)The problem is solved by generating multiple hypotheses about the fraction of valid data and then checking error model conformity to determine the best hypothesis. Assuming that the computational cost of the error model conformity tests is negligible in comparison with the optimisation process aimed at estimating the parameters, the overall cost is proportional to the number of hypotheses tested.
6.4 Tracking results

The new shape-constrained tracker was tested on the 5 shots of the M2VTS database [57] (see chapter 4). Tracking was again stable for all the subjects, but in this case with clear improvements in shape accuracy, with lip contours that now look much more natural and are significantly less noisy. This can be seen by comparing figures 6.1 (every tenth frame of the sequence) and 6.2 (zooming on the mouth area) with the corresponding pictures (figures 4.10 and 4.11) shown in chapter 4 for the same example sequence ('sp.04.v').

Although at a significantly lower scale, tracking experiments were also performed on sequences of an extended M2VTS database, with similar results.

Most of the tracking inaccuracies can be attributed to the mismatch between the colour characteristics of the lips (see results for frames 90-94 in Figure 6.2 above) and those of the built colour models. Let us recall that colour models were computed at once from the first image of the sequence, without further updates, and that lip colour can be expected to show a certain range of variation during speech as a consequence of the varying flow of blood underneath the lip epidermis [140] during the various phases of speech production. Nevertheless, the problem is not so serious as with the bootstrap tracker because now the existence of a shape model and its prior distribution favours plausible shapes supported by the most reliable measurements.

Tracking is actually lost by the end of some sequences when the subject, upon finishing speech utterance, rapidly turns his/her head, resulting in very quick horizontal lip motion that the tracker cannot follow. Other than that, the new tracker has demonstrated that it can cope with significant facial motion during some sequences, although some temporary inaccuracies can be noticed for some frames before the tracker fully recovers. In these cases, the optimum of the criterion function also shows values higher than expected, because of also higher spatial discrepancy between the prediction and the current measurements.

More sophisticated dynamics models could then help further improve tracking quality by generating better predictions upon which the new measurements would be extracted. It is worth stressing, anyway, that these disturbing facial movements are not really characteristics of speech production, and rather correspond to spontaneous gestures or mannerisms. As opposed to creating complex dynamics models that take all these factors into account, Reynard et al [135] suggested the use of coupled trackers: a primary tracker would track the face and a second one, coupled to the first, would keep track of the lips.

6.5 More complex dynamics models?

There are Finite Element Method approaches that claim the use of 'physically'-based constraints for object dynamics. However, despite what the term may suggest, they typically have little to do with the real physics underneath the perceived phenomena. Taking into account the action of all muscles involved in mouth movement,
Figure 6.1: sp.04.v: tracking results for every tenth frame, in raster order
6.5. More complex dynamics models?

Figure 6.2: sp_04.v: tracking results for every frame (0-112), in raster order
as well as the special characteristics of the lips as deformable, non-rigid objects, and eventually considering how this all impacts on image forming\(^2\), is a daunting task. Not surprisingly, as can be seen in [118], the application of those so-called ‘physically’-based constraints is not an attempt to construct a physiological model having a simple relation to the actual stiffness of the skin, muscle and other tissue that make up the mouth region. What is actually done is modelling the visible observations of the mouth.

Kalman filtering\([158]\), is another well known technique widely used for tracking purposes (e.g. [112]). The Kalman filter is optimal with a correct dynamics model and gaussian (monomodal) probability distributions. Tracking behaviour can be quite poor with wrong dynamics modelling [124], let alone the effect of multimodal distributions [159]. In [13], the actual dynamics are learnt from sequences like the ones that would be tracked thereafter. This involves the use of some sort of autoregressive model that recovers the expected dynamics of the modelled objects. As a result, tracking performance is enhanced through the construction of rather specific object dynamics models.

In [113] lip dynamics is represented by a (constant) second order autoregressive model learnt from a training sequence \(\{\theta_1, \ldots, \theta_M\}\):

\[
\theta_n = A_2 \theta_{n-2} + A_1 \theta_{n-1} + D
\]  

(6.7)

where \(A_2, A_1,\) and \(D\) are estimated as follows:

1. Compute sums \(R_i\) and autocorrelation coefficients \(R_{ij}, R'_{ij}\) \((i, j = 0, 1, 2)\):

\[
R_i = \sum_{n=3}^{M} \theta_{n-i}
\]

\[
R_{ij} = \sum_{n=3}^{M} \theta_{n-i} \theta_{n-j}^T
\]

\[
R'_{ij} = R_{ij} - \frac{1}{M-2} R_i R_j^T
\]

(6.8)

2. The estimated parameters \(\hat{A}_2, \hat{A}_1,\) and \(\hat{D}\) are then given by:

\[
\hat{A}_2 = (R'_{02} - R'_{01} R'^{-1}_{11} R'_{12}) (R'_{22} - R'_{21} R'^{-1}_{11} R'_{12})^{-1}
\]

\[
\hat{A}_1 = (R'_{01} - \hat{A}_2 R'_{21}) R'^{-1}_{11}
\]

\[
\hat{D} = \frac{1}{M-2} (R_0 - \hat{A}_2 R_2 - \hat{A}_1 R_1)
\]

(6.9)

In order to check how advantageous this model could be, the prediction error was measured for a whole tracking sequence of frames and compared with that obtained using a first-order regressive model consisting of taking the last measurements as the current state prediction. The results (see Table 6.1) show just moderate improvements when the more complicated second order model is used.

\(^2\)variable lip reflectivity, number and spectral characteristics of the illuminants, pose, obscurcation, etc
6.5. More complex dynamics models?

<table>
<thead>
<tr>
<th></th>
<th>1st order</th>
<th>2nd order</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>1.659</td>
<td>0.583</td>
</tr>
<tr>
<td>median</td>
<td>0.727</td>
<td>0.439</td>
</tr>
</tbody>
</table>

Table 6.1: Mean and median squared prediction errors in the 'sp.01_v' sequence. The autoregressive 2nd order model achieves slight prediction improvements.

The point about using a second order autoregressive model is that it is the simplest model that can account for both translational and oscillatory motion. Blake and Isard [113] point to the pseudoperiodicity of lip motion during speech, although it appears that distinct periodic elements occur over a spread of frequencies. We can see this in Figure 6.3. Figure 6.3(a) shows the evolution of the vertical scaling parameter and the coefficient associated with the first eigenlip as functions of time. In Figure 6.3(b), the spectra of both signals are presented. We can see how both scaling (it is also represented after subtracting the average, to illustrate the non-DC component) and the shape parameter show quite oscillatory behaviours, and there are no clear resonant peaks - which would very much support the benefit of a second order model - but rather a number of them.

These results would suggest that using all frames of a sequence to estimate the parameters can provide too coarse a representation that does not properly reflect short term correlation detail, as opposed to what happens, for instance, in speech analysis, where linear predictive coefficients prove a very satisfactory means to capture the short term characteristics of the speech signal. Similarly, as done in the speech domain, we could think of allowing for time-varying adaptive autoregressive models (e.g. [160]), i.e. whereby the parameters of the AR model are updated to account for variations of the signal characteristics over time. In practice, this is usually done...
Table 6.2: Mean and median squared prediction errors in the ‘fm_01.v’ sequence using ‘sp_01.v’ models. Gross prediction errors are obtained when using the AR model computed for a different subject/sequence.

<table>
<thead>
<tr>
<th></th>
<th>1st order</th>
<th>2nd order</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>5.000</td>
<td>182.554</td>
</tr>
<tr>
<td>median</td>
<td>2.817</td>
<td>106.553</td>
</tr>
</tbody>
</table>

by restricting the estimation of the model parameters to a sliding window consisting of a predetermined number of samples (e.g. 10-th order LPC estimations are typically performed on 20 ms frames, i.e. 160 samples at 8 KHz sampling rate) where the mechanisms of speech production are supposed stationary.

Unfortunately, there does not seem to be much room to capture short term correlation at its best. The tracked sequences correspond to the utterance of digits 0 to 9 in French. This means that we have of the order of a few tens of phonemes or diphones, and of the order of a hundred of frames. Even if we restrict the estimation of the autoregressive model to a number of frames spanning approximately the duration of a phoneme or part of it, the number of frames per estimation we would be left with would result in numerical problems (autocorrelation matrices becoming rank deficient) and in little statistical significance. This point is actually made in [161]: ordinary least squares and maximum likelihood methods are ruled out, and Kalman filtering is proposed to perform an on-line estimation of the time-varying autoregressive parameters, just assuming a simple ‘dynamics’ model (such as random walk) for the regression parameters.

Another characteristic of the ‘constant’ autoregressive model learnt from a certain training sequence is its specificity. To verify to what extent this happens, the second-order AR model learnt with a sequence corresponding to a given speaker was used to fit the tracking results obtained with a different sequence corresponding to another speaker. In this case, the first-order model yielded clearly better results than the learnt second-order model (Table 6.2).

Specificity is still compatible with a personal verification scheme where each subject whose identity is to be verified would have a specific dynamics model. When the claimed identity is correct, tracking would be expected to work reliably so that lip-based features could further contribute to the reliability of the verification system.

6.6 Discussion

Although tracking performance shows significant countour accuracy, colour imprecisions still result in occasional local errors where the actual lip boundary is missed by a couple of pixels. Three strategies could be implemented to further improve tracking performance. A first one stems from better colour modelling: robust clustering techniques [152] would be applicable to estimate better class-dependent colour
6.7. Summary

A robust, shape-constrained lip contour tracking has been described. A linear combination of eigenlips is affine-warped to achieve the lowest reconstruction error of a number of lip boundary estimates drawn from the current estimation.

Lip dynamics is represented by a simple first-order model where the current estimate is used as an a priori prediction for the following frame. A Dynamic Hill Climbing models. This strategy has actually been implemented (see Appendix A) and successfully tested in the XM2VTS database [9] (see chapter 7), resulting in better colour estimation, and consequently in smoother, more accurate lip contour tracking. A second strategy could focus on the process of profile sampling and lip boundary estimation. It is common to model the lip contour as a boundary between two regions, namely, the lips and the surrounding area. This brings along the inconvenience of having to establish some workable assumptions about that surrounding area, typically that it corresponds to skin. Whereas this is customarily the case, subjects with beards—which is certainly not uncommon—pose a problem. Therefore, it would be advantageous to carry out boundary estimation through determining the end of the compliance with a lips hypothesis, without regard to an external surrounding area. Finally, a third strategy could consist of adapting colour models during tracking [116, 162]. Moreover, provided the other ambient conditions (mainly lighting) remain constant, the actual colour drift would potentially be very relevant for reflecting physical changes in the lips associated with speech production.

Currently, lip tracking is biased towards average-like shapes because of the probabilistic shape weighting imposed (Mahalanobis distance term in Equation 6.6). Another possibility is to allow for as much shape variability as reasonably acceptable: we could apply the shape correction term only when the eigenlip components exceed a certain number of standard deviations—thus penalising odd shapes—and keep a constant term otherwise.

The present tracking scheme relies on determining a measure of reliability for each extracted measurement and then assigning it a probability-based weighting. An alternative could be the investigation of the CONDENSATION algorithm [133, 159] in this context. The CONDENSATION algorithm specifically accommodates the possibility of occlusion or failure to detect features in an attempt to track curves in dense visual clutter. In these conditions, the Kalman filter is inadequate and, because of being based on unimodal Gaussian densities, cannot represent simultaneous alternative hypotheses. In the CONDENSATION algorithm, the probability distribution of possible interpretations is represented by a randomly generated set which is propagated over time according to a learned dynamics model and the visual observations. Once again, we are faced with the problem of determining valid dynamics and observational models. Furthermore, factored sampling can be inefficient as the models of the conditional observations probability density function become narrow.

6.7 Summary
algorithm is used for finding both the eigenlip coefficients and the parameters of the affine transform as the solutions of the reconstruction error minimisation problem.

Tracking experiments in the M2VTS show stable tracking, with good contour accuracy in general, except for some local inaccuracies as a result of poor colour characteristics, or sudden non speech-related facial motion. Although more complex dynamics models are deemed advantageous, the simple first order model seems to be satisfactory enough for dealing with lip motion during speech production.
Chapter 7

Application of lip-based features to the verification of personal identity

7.1 Introduction

References to lip-reading applications for speech recognition and synthesis are abundant in the literature but, other than the work done by Lüttin [101] or Mason et al. (e.g. [136]), there does not seem to be that much research on lip-based speaker verification (recognition). However, as suggested in [163], even very coarse lip features can be used as behavioural biometric characterisation of the speaker or as a means for detecting the lip shape status which in turn can serve as a control information for face coding or recognition. The latter was demonstrated in [164], where a B-spline lip tracking system like the bootstrap tracker of chapter 4 was used to provide control information regarding the state of the lip shape which is used by a conventional eigenface-based face verification system to confirm or reject a claimed personal identity. The performance of the system tested on the M2VTS database [57] showed a promising improvement over the unimodal approach. This improvement derives from the achieved reduction in the population entropy of the models, thus minimising the probability of impostor acceptance.

In this chapter it will be shown how the information supplied by the shape-constrained lip tracker of chapter 6 can be advantageously used to implement a text-dependent speaker verification system based exclusively on lip shape features. This verification modality is thereafter combined with other visual and vocal experts, resulting in improved overall performance. The experiments reported in this section were carried out in the XM2VTS database [9] according to the Lausanne Protocol [10]. Although these results on the application of the lip tracker can be considered preliminary, they are already quite illustrative of the kind of improvements to expect.
7.2 Description of the XM2VTS database and the Lausanne protocol

The XM2VTS database contains synchronised image and speech data as well as sequences with views of rotating heads. The database includes recordings of 295 subjects taken at one month intervals. In each session two recordings were made, each one consisting of a speech shot and a head rotation shot. The speech shot consisted of a frontal face recording of each subject during speech production, namely the utterance of three speaking sequences: two digit sequences, and a sentence.

The Lausanne protocol is a published evaluation proposal for the XM2VTS database. Two protocol configurations were defined (see the corresponding database partitionings in Tables 7.1 and 7.2):

- Configuration I: The assumption is good expert training using data from three different sessions, and inferior fusion training using data from the same shots that were used for expert training
- Configuration II: The assumption is inferior expert training using data from only two different sessions, and good fusion training using data from shots that were not used for expert training

Each shot being used consists of the 2 audio digit sequences and of one image. The 295 subjects were divided into three sets: 200 clients, 25 impostors for evaluation, and 70 impostors for independent testing. The impostors in the evaluation set allow to train a supervisor with impostors that were never seen by the experts. The evaluation set serves for the evaluation of experts, the determination of the verification threshold, and for the training of the supervisor.

This leads to the following statistics:

- Client training examples: 3 per client in Configuration I, 4 per client in Configuration II.
- Evaluation samples (clients): 600 in Configuration I, 400 in Configuration II.
- Evaluation samples (impostors): 40000 \((25 \times 4 \times 2 \times 200)\).
- Test client accesses: 400 \((200 \times 2)\).
- Test impostor accesses: 112000 \((70 \times 4 \times 2 \times 200)\).

7.3 Extraction of lip features and matching strategy

The lip tracker described in the previous chapter was used to extract lip features from the two audio digit sequences available in each shot. In the current text dependent
### Table 7.1: Partitioning of the XM2VTS database according to Configuration I

<table>
<thead>
<tr>
<th>Session</th>
<th>Shot</th>
<th>Clients</th>
<th>Impostors</th>
<th>Impostors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Training Data</td>
<td>Evaluation Data</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Evaluation Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Training Data</td>
<td>Evaluation Data</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Evaluation Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Training Data</td>
<td>Evaluation Data</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Evaluation Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Test Data</td>
<td>Evaluation Data</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 7.2: Partitioning of the XM2VTS database according to Configuration II

<table>
<thead>
<tr>
<th>Session</th>
<th>Shot</th>
<th>Clients</th>
<th>Impostors</th>
<th>Impostors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Training Data</td>
<td>Evaluation Data</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Evaluation Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Training Data</td>
<td>Evaluation Data</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Evaluation Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Evaluation Data</td>
<td>Test Data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Test Data</td>
<td>Evaluation Data</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
lip-based verification that will be described below, the two audio sequences are concatenated, resulting in a single, bigger sequence.

Colour models for each client were computed off-line using the robust clustering algorithm described in Appendix A. For each subject, color models were extracted from 4 frontal images of the head rotation set, each from a different session, and then averaged to generate an all-session subject-specific colour model. In those sessions where the colour characteristics significantly differed from the model (for instance, due to the presence/absence of lipstick or to dramatic changes of lipstick color preferences in some of the cases where it was worn), session-specific colour models were used instead. Lip tracking initialisation was performed using the procedure described in Appendix B.

As far as tracking performance itself is concerned, its importance is very much acknowledged, but lacking an objective and meaningful quality metric, such an analysis will be omitted and, as previously pointed out by Jourlin et al [165], it is the combined performance of tracking and verification that will be evaluated through the verification experiments that will be described in the coming sections. Nonetheless, the subjective impression is quite good for most of the speakers, which is quite remarkable in view of the broad ethnical background coverage of the XM2VTS, and the fact of having to cope occasionally with significant motion. The results also show the good generalisation capability of the eigenlips estimated from the M2VTS database (see chapter 5). Problems have, however, been detected a) with some speakers wearing dark beards and/or moustaches, b) with speakers where lip colour is hardly distinguishable from the surrounding skin, specially if relatively reddish areas occur in the surrounding skin area, and c) unusual degree of motion. The presence of surrounding skin together with moustaches/beards already violates the working assumption of estimating the lip contour as the boundary between two homogeneous regions, with the added difficulty in these cases that the estimation of a unimodal colour model of the area surrounding the lips (by merging quite different chromaticity clusters results into a single one) results in a model which happens to be closer to the lips model than any of the constituting clusters. In case a), although tracking remained stable it failed to accurately follow the lip contour outline. Case b) was less severe, and tracking failure was restricted more often to temporary distractions, typically involving the lower contour of the lips. As far as case c) is concerned, even significant degrees of motion were generally well tolerated, although in a few cases allowing for a temporary, partial loss of tracking (for instance affecting a corner of the mouth), prior to recovery.

The lip tracker supplies a set of eigenlip coefficients and affine transform parameters for each frame. By warping the linear combination of eigenlips with the affine transform parameters, a 22-dimensional feature vector is obtained that consists of the geometrical coordinates of the 11 control points used for characterising the lip contour. Accordingly, an utterance consisting of \( N \) frames is represented by a sequence of control point vectors \( \mathbf{u}_1, \ldots, \mathbf{u}_N \) which define a trajectory in a 22-dimensional space. Verification tests are operated by matching the trajectory under test \( T = \{ \mathbf{u}_1, \ldots, \mathbf{u}_N \} \) against a reference template \( R = \{ \mathbf{v}_1, \ldots, \mathbf{v}_M \} \) corresponding to the claimed identity using a Dynamic Time Warping Algorithm (DTW) [166].
A framewise dissimilarity metric is given by

\[ d(u_i, v_j) = (u_i - v_j)^T H (u_i - v_j) \]  

(7.1)

where \( H \) is the metric matrix that was introduced in chapter 5, and which converts control point distances into real shape distances.

Unless additional information becomes available, it is generally not possible to establish whether affine variation just corresponds to pose variations (e.g. different lips size due to posing at a different distance from the camera) or they are really characteristic of a given identity. Furthermore, even small, perceptually irrelevant changes in scale can have a stronger impact on the metric considered than proper shape variation. This is why eventually all control point vectors are normalised for translation, scale and rotation. Hence the control point constellation for each frame is translated to the origin of coordinates and the point set is rotated so that the points corresponding to the mouth corners on the B-spline contour are aligned with the \( x \) axis. Finally, to account for scale normalisation, whilst allowing for relative size changes in the sequence of frames corresponding to a single utterance, the mouth width mode \( w_0 = \text{mod}_{i=1}^{N} w_i \) is computed in a first pass and then the control point constellation of each frame is homogeneously (both horizontally and vertically) scaled by a factor \( \frac{w_{ref}}{w_0} \) in a second pass, where \( w_{ref} \) is a predefined width value. As a result of this process, the mouth shapes of every single utterance will have their most common width set to this reference value.

### 7.4 Lip-based verification of personal identity

To provide the present DTW-based text-dependent speaker verification scheme with statistical foundations, the DTW distance distributions of client and impostor populations of the XM2VTS database were analysed [150]. So as to decouple this analysis as much as possible from the realisation of the verification experiments according to the Lausanne protocol, the impostor population was built from the actual client set, i.e. for each of the 200 'clients' of the XM2VTS database, the other 199 are considered as 'impostors'.

Data from shots 1 and 2 of sessions 1 and 2 (this corresponds to the training data according the Configuration II of the Lausanne protocol) was used for generating the reference templates for each of the 200 clients, and data from shots 1 and 2 of session 3 for evaluation purposes. Thus, 4 sequences were available for each speaker for training, and 400 (2 \( \times \) 200) sequences were left for evaluation.

Two different approaches were considered:

- Standard procedure: training proceeds by successively aligning and averaging into a reference template each of the 4 utterances available per subject. The matching score \( D \) during evaluation is given by the average distance over the optimal path.
Chapter 7. Application of lip-based features to the verification of personal identity

Table 7.3: Equal error rates (EER) and associated thresholds for the standard and robust procedures considered.

<table>
<thead>
<tr>
<th></th>
<th>EER</th>
<th>threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>17.7%</td>
<td>41.8</td>
</tr>
<tr>
<td>robust</td>
<td>15.8%</td>
<td>28.5</td>
</tr>
</tbody>
</table>

- Robust procedure: the reference template is built from the most compact subset of 3 utterances selected among the 4 available per speaker. The remaining utterance is not rejected but kept as a fall-back template. Therefore, two matching scores are obtained in this case, the minimum of the two being finally selected. As another difference with reference to the standard procedure, each of the scores is given by the median distance over the optimal path.

The second approach aims at reducing the effect of occasional gross errors on training by supporting utterance consistence, and by disregarding temporary gross differences (as it might happen, for instance, if tracking quality degrades temporarily for a few frames) through the use of a median-based score.

Figure 7.1 shows the distance probability density functions (PDF) and the corresponding receiver operating characteristics (ROC) for the two procedures considered. False rejection is the case where a client, claiming his/her true identity, is rejected. False acceptance is the case where an impostor, claiming the identity of a client, is accepted. The PDFs are estimated through the normalised histograms of the available evaluation distance scores: 400 client samples (2 x 200), and 79600 impostor samples (2 x 199 x 200). Assuming verification tests are operated by comparing the measured distance against a global threshold $T$, the relationship between distance PDFs and ROCs becomes evident. If we call $\psi_C(D)$ and $\psi_I(D)$ the respective PDFs of clients and impostors, the false rejection $FR(\tau)$ and false acceptance $FA(\tau)$ characteristics can be computed as follows [150]:

$$FR(\tau) = \int_{-\infty}^{\tau} \psi_C(D)dD = 1 - \int_{0}^{\tau} \psi_C(D)dD$$
$$FA(\tau) = \int_{\tau}^{\infty} \psi_I(D)dD$$

(7.2)

The equal error rate (EER) threshold is determined by the intersection of the $FR(\tau)$ and $FA(\tau)$ characteristics. The results are shown in Table 7.3. As can already be seen from the graphics, the robust procedure results in more compact distance scores, specially for the clients. Consequently, the EER threshold is lower than in the standard case and, more importantly, so is the EER proper.

The Lausanne protocol focuses on verification. However, as a byproduct of the previous analysis, recognition performance was measured as well. Each of the 400 evaluation sequences was matched against each of the 200 client templates and classified accordingly using a nearest neighbour rule. The correct classification rates are shown in Table 7.4.
7.4. Lip-based verification of personal identity

Figure 7.1: Distance probability density functions (PDF) and corresponding receiver operating characteristics (ROC) for the standard and robust procedures, respectively.

<table>
<thead>
<tr>
<th></th>
<th>1st shot</th>
<th>2nd shot</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>55.50 %</td>
<td>57.00 %</td>
<td>56.25 %</td>
</tr>
<tr>
<td>robust</td>
<td>62.50 %</td>
<td>62.50 %</td>
<td>62.50 %</td>
</tr>
</tbody>
</table>

Table 7.4: Speaker recognition rates (experimental data from shots 1 and 2 of session 3)
Although recognition performances per se are admittedly moderate, they are still two orders of magnitude higher than chance performance and the figures compare pretty well with the obtained by Luttin [101] in the smaller M2VTS database (37 speakers) when only shape information is considered\(^1\) (about 53% recognition rate using shape features only and 60% when first order temporal difference parameters are also included).

### 7.5 Speaker verification results on the XM2VTS database according to the Lausanne protocol

The robust procedure described in the previous section was adopted for carrying out the experiments. The ROC curve of Figure 7.1 was used to map the distance scores generated by the DTW routine into the interval \([0,1]\) through the following transformation:

\[
D^*_0(D) = \frac{1 + FR(D) - FA(D)}{2}
\]  

(7.3)

By construction, \(D^*_0(\tau) = 0.5\). For small distances, \(D^*_0\) tends asymptotically to 1; conversely, for large distances \(D^*_0\) tends to 0.

In Configuration I the robust training strategy builds the reference template for each client from the two nearest utterances out of the set of 3 available for training, and keeps the third one as a fall-back template which is also used for matching during evaluation and testing. The generation of reference templates according to Configuration II was already described in the previous section.

Global thresholds for verification experiments in each configuration are estimated on the corresponding evaluation sets. These thresholds, together with the final verification results (false rejection rate FRR and false acceptance rate FAR) obtained after independent testing\(^2\) can be seen in Table 7.5. The error rate is about 14% on average, quite close to what could be expected from the statistical analysis of the statistical distance distribution that was presented in the preceding section.

### 7.6 Fusion experiments on the XM2VTS database according to the Lausanne protocol

The combination of a number of experts can potentially improve - and sometimes significantly - the performance attained by the best individual modality. This was

---

\(^1\)Recognition rates of about 80% and higher were reported when the shape features were augmented with intensity information, and the overall best performance was actually attained using only intensity features.

\(^2\)The recording of two video sequences of the impostor test set was found faulty and, therefore, the reported verification results are based on 111600 impostor accesses and not on the theoretical 112000. This circumstance was also taken into account during the fusion experiments, and the results corresponding to these two sequences were excluded.
7.6. Fusion experiments on the XM2VTS database according to the Lausanne protocol

<table>
<thead>
<tr>
<th>Configuration</th>
<th>threshold</th>
<th>FRR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.50</td>
<td>14.00 %</td>
<td>12.67 %</td>
</tr>
<tr>
<td>II</td>
<td>0.46</td>
<td>11.75 %</td>
<td>17.33 %</td>
</tr>
</tbody>
</table>

Table 7.5: Lip-based speaker verification results

the theme of the M2VTS project [2], and more recently, also in the context of the XM2VTS database and the Lausanne protocol, successful integration result were reported in [11].

The fundamentals underpinning fusion is that by drawing on several independent sources of information, an adequate combination of them can overcome the shortcomings and limitations of each of the individual modalities. The converse is also possible (performance degradation) and from that it follows the importance of developing appropriate information integration strategies.

Typical fusion strategies consist of simple combination rules: maximum, minimum, median, average score, and product of scores. Conditions under which such schemes perform well are theoretically understood and have been shown to hold in applications [167]. However, in a very similar fusion scenario [11] (in fact some of the experts combined are also used here) combining high performance speech verification modules and a medium vision module (face recognition), the conditions were violated and none of the aforementioned fusion schemes performed better than the best individual expert.

In the light of those considerations, and the successful performance obtained with a linear weighted combination rule [165], this was eventually the fusion strategy adopted for these experiments. According to this integration paradigm, a verification score \( v \) is obtained as a linear combination of the scores of the \( m \) modalities to fuse (\( v = w_1 v_1 + \ldots + w_m v_m \)) and then compared with a threshold \( \tau_0 \). The optimal weights \( w_1 \ldots w_m \) and the acceptance threshold \( \tau_0 \) are chosen using the evaluation set.

Apart from the described DTW-based text dependent verification system based on lip features (SURREYL), 3 face recognition algorithms (SURREY1 -based on robust correlation [56] - SURREY2 -Linear Discriminant Analysis [73] - and AUT1), and 2 voice-based modalities [11] (IDIAP2 -sphericity- and IDIAP3 -HMMs) were considered for the fusion experiments. Their individual performances (Configuration I) on the test set are shown in Table 7.6.

The following fusion experiments were considered:

1. Lips and face (SURREY2)
2. Lips and voice (IDIAP3)
3. Face (SURREY2) and voice (IDIAP3)
4. Lips, face (SURREY2) and voice (IDIAP3)
Chapter 7. Application of lip-based features to the verification of personal identity

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>threshold</th>
<th>FRR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURREY1 (lips)</td>
<td>0.50</td>
<td>14.00 %</td>
<td>12.67 %</td>
</tr>
<tr>
<td>SURREY1 (face)</td>
<td>0.50</td>
<td>7.25 %</td>
<td>7.78 %</td>
</tr>
<tr>
<td>SURREY2 (face)</td>
<td>0.21</td>
<td>5.00 %</td>
<td>4.45 %</td>
</tr>
<tr>
<td>IDIAP2 (voice)</td>
<td>0.50</td>
<td>7.00 %</td>
<td>1.42 %</td>
</tr>
<tr>
<td>IDIAP3 (voice)</td>
<td>0.50</td>
<td>0.00 %</td>
<td>1.48 %</td>
</tr>
<tr>
<td>AUT1 (face)</td>
<td>0.50</td>
<td>6.00 %</td>
<td>8.12 %</td>
</tr>
</tbody>
</table>

Table 7.6: Performance of modalities on test set (Configuration I).

<table>
<thead>
<tr>
<th>Modalities</th>
<th>weights</th>
<th>threshold</th>
<th>FRR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>lips and face</td>
<td>0.42, 0.58</td>
<td>0.38</td>
<td>4.50 %</td>
<td>0.73 %</td>
</tr>
<tr>
<td>lips and voice</td>
<td>0.41, 0.59</td>
<td>0.54</td>
<td>0.00 %</td>
<td>1.39 %</td>
</tr>
<tr>
<td>face and voice</td>
<td>0.58, 0.42</td>
<td>0.42</td>
<td>0.00 %</td>
<td>1.25 %</td>
</tr>
<tr>
<td>lips, face and voice</td>
<td>0.27, 0.23, 0.49</td>
<td>0.51</td>
<td>0.00 %</td>
<td>1.31 %</td>
</tr>
<tr>
<td>5 modalities (no lips)</td>
<td>0.00, 0.02, 0.87, 0.05, 0.06</td>
<td>0.50</td>
<td>0.00 %</td>
<td>0.52 %</td>
</tr>
<tr>
<td>all 6 modalities</td>
<td>0.03, 0.00, 0.01, 0.89, 0.03, 0.04</td>
<td>0.50</td>
<td>0.00 %</td>
<td>0.29 %</td>
</tr>
</tbody>
</table>

Table 7.7: Fusion results (Configuration I).

5. All: SURREY1, SURREY2, IDIAP2, IDIAP3 and AUT1

6. All: SURREY1, SURREY2, SURREY1, IDIAP2, IDIAP3 and AUT1

The results, as well as the corresponding optimal combination weights and acceptance threshold can be seen in Table 7.7. It is interesting to see how in all cases the trained linear weighted classifier performs better than the best individual expert. It is also worth remarking how the 4th fusion strategy (lips, face and voice) does perform slightly worse than the 3rd one (face and voice), which can be put down to overtraining since the former did yield a lower FAR figure during evaluation. Eventually the best results among the 6 test scenarios considered are obtained when all 6 modalities are combined, although it can be seen how the weights attributed to some of them are quite low, or even zero. In order to see to what extent lip features do represent a positive contribution to the overall performance, the results for a trained classifier combining the other 5 modalities, leaving aside the lips, are shown as well. Getting further improvements at low error rates is very difficult and lips reduce the error rate of the 5 modality case by roughly 40%.

In Configuration II, fusion experiments could also be performed using face recognition data (SURREY2) in combination with lip information. The individual performances of each modality are shown in Table 7.8, and the fusion results in Table 7.9. As it happened with the experiments according to Configuration I, there is also a clear gain by combining both classifiers, resulting in a significant reduction of the false acceptance rate. It is worth mentioning how -in comparison with the corresponding experiments in Configuration I, more weight is now given to the face
7.7. Conclusions

A text-dependent DTW-based person identity verification system using lip features during speech production has been presented. The system builds upon the tracking results generated by a shape-constrained chromaticity-based lip tracker which was run for the more than two thousand audio sequences of the XM2VTS database.

The verification performance of this lip-based modality was tested according to the Lausanne protocol, with error rates of about 14% on average in both configurations. Moderate though (in comparison with other verification modalities also tested on the XM2VTS database), it is worth stressing that these figures are reflecting the combined performance of tracking and the discriminatory information provided by the outer lip contour. It is therefore expected that these error figures can be further reduced through tracking improvements, but also by making use of feature selection techniques [166, 168], or by augmenting the shape information with additional information (e.g. additional shape information, temporal variation, or intensity [101, 165]).

More importantly, it has been demonstrated, how a 'weak' verification modality brings in additional discriminatory information that can result in improved overall performance when combined with other verification experts. Experiments carried out with a trained weighted linear classifier combining different verification modalities (face, voice, lips) showed, in all cases considered, better verification performance than the best individual modality being combined.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>threshold</th>
<th>FRR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURREY1 (lips)</td>
<td>0.46</td>
<td>11.75 %</td>
<td>17.33 %</td>
</tr>
<tr>
<td>SURREY2 (face)</td>
<td>0.25</td>
<td>1.00 %</td>
<td>2.07 %</td>
</tr>
</tbody>
</table>

Table 7.8: Performance of modalities on test set (Configuration II).

<table>
<thead>
<tr>
<th>Modalities</th>
<th>weights</th>
<th>threshold</th>
<th>FRR</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>lips and face</td>
<td>0.25, 0.75</td>
<td>0.32</td>
<td>1.00 %</td>
<td>0.63 %</td>
</tr>
</tbody>
</table>

Table 7.9: Fusion results (Configuration II).

expert, the performance of which is better in Configuration II than in Configuration I, whereas the opposite happens as regard the lip expert.
Chapter 7. Application of lip-based features to the verification of personal identity
Chapter 8

Conclusions

8.1 Research contributions

It has been shown how a statistical characterisation of the chromaticity values of both the lips and the surrounding skin can be exploited to construct a lip boundary tracker (chapter 4) without hardly any other supporting constraints but the smoothness and the relatively reduced number of degrees of freedom of a B-spline representation (e.g. [121, 137]). The tracker extracts profiles normal to the current estimate of the lip boundary and generates new boundary estimates by taking into account in a statistical framework both the spatial context and the colour characteristics. A new B-spline is computed as the most-likely fit of the extracted measurements, which admits a weighted least squares formulation where the weights are probability terms associated with each of the lip boundary measurements.

The so-called bootstrap tracker is then advantageously used to construct a rather accurate and robust shape-constrained eigenlip tracker (chapter 6) also based on the above statistical chromaticity characterisation and where shape variation is limited to affine deformations of a linear combination of a reduced number of principal modes of variation. The thesis shows how the main statistical models of lip-shape variation can be computed from the results obtained with the aforementioned shape-unconstrained bootstrap tracker (chapter 5). Although colour information is in itself insufficient and the tracker may yield a considerable number of wrong lip shapes, a set of representative eigenlips can be computed through a robust estimation of the lip shape covariance matrix under the assumption that there is still a majority of valid tracked shapes.

The estimation exploits a novel adaptation of Rousseeuw's MCD method [16] characterised by its high breakdown point and its good statistical convergence. Similarly to [101] tracking dynamics follows a simple first-order model where the estimate obtained for the preceding frame is used as the prediction for the current one. The current estimation is obtained through minimisation of the probabilistically-weighted reconstruction error of the new lip boundary estimates. As opposed to the approach followed in the bootstrap tracker, the variability is restricted to a reduced number of principal modes of shape variation which are affine transformed to match the
measurements. A Dynamic Hill Climbing algorithm [156] was used to estimate the optimal shape and affine-warp parameters. Although the objective function is not robust in strict sense because of its low breakdown point [151], the probabilistic weighting of the measurements ensures that the role played by each measurement is commensurate to its goodness.

The estimation of colour models is of central importance for satisfactory tracking performance. Thus the ISODATA-like clustering routine used for the bootstrap tracker gave way to the implementation of robust chromaticity clustering algorithms (Appendix A) that the eigenliptracker could take advantage of. The optimal number of clusters is also automatically estimated as part of the process according to the goodness of fit of the clusters to the overall colour measurements. A first ‘crisp’ version was implemented based on the Generalised Minimum Volume Ellipsoid method described in [169] and Rousseeuw’s MCD estimator [16]. To cater for cases where cluster separability is not that good, an alternative fuzzy version was implemented, whereby cluster membership functions are defined through the consideration of class (cluster) conditional probability distributions.

The eigenlip tracking system was successfully tested on the M2VTS database, and more recently on the more than 2000 video sequences (each typically consisting of more than 300 frames) of the XM2VTS database in order to extract visual features for personal identity verification. The overall impression is good, specially when taking into account the broad ethnical coverage of the database, and the challenging conditions posed by some subjects with beards and moustaches.

Eventually, verification experiments were performed on the XM2VTS database according to the Lausanne protocol (chapter 7) using the tracking results supplied by the aforementioned eigenliptracker. The lip features consisted of the B-spline control points associated with each frame, once normalised for translation, rotation and scale, whilst allowing for relative size changes during a given tracked utterance. A DTW algorithm was used for computing a distance score between a test sequence and a template corresponding to the claimed identity. Measured average error rates are approximately 14\% in both protocol configurations. Although this performance is inferior to that of other traditional approaches like face or voice, it shows that the variation of the shape of the outer contour of the lips during speech production also contains a significant amount of discriminatory information.

Finally, a weighted linear classifier combining different verification modalities together with the lip-based system was trained on the evaluation set for information fusion tests. In all cases considered, the combined multimodal system outperformed the best individual expert, thus showing the power of combining information coming from independent sources and that even theoretically ‘weak’ verification modalities (like the lip-based modality in comparison with face or voice systems) can make a contribution, sometimes significant, towards a better overall performance.
8.2 Future work

Some ideas for further lip tracking improvements were already suggested in chapter 6, namely the on-line adaptation of colour models, and the revision of the two-class lip boundary estimation. Colour model adaptation was originally suggested to account for changes and inaccuracies of the original model. It could also be exploited at the tracking initialisation stage (currently depending on the existence of precomputed colour models), and for augmenting the shape features extracted by the tracker with information about colour variation, some of which could be characteristic of person identity. As regards lip boundary estimation, the use of multimodal colour models for each modelled region appears as a straightforward extension.

In addition to that, a better dynamics model is expected to offer a twofold benefit: better prediction propagation than with the simple first-order model, and more appropriate constraints for affine deformation across frames (whereas the eigenlips parameters can be understood to follow a certain probability distribution that limits in practice the values a shape can adopt, no such constraints apply on the parameters of the affine transform, other than the need to minimise an error function). As it was also suggested in chapter 6, the use of adaptive autoregressive models -already applied in other domains [160, 161] where the stationarity characteristics of the signal change over time- could be advantageous in this context.

As far as lip-based verification is concerned, the visual features being currently used are likely to be significantly correlated and, therefore, feature selection techniques are potentially very interesting to accomplish both dimensionality reduction and a better verification performance, as was already demonstrated by Pandit et al [166, 168] in a similar context.
Chapter 8. Conclusions
Appendix A

Robust clustering

A.1 Introduction

A robust clustering algorithm based on the minimum covariance determinant (MCD) estimator [16] that was described in chapter 5 is proposed. In the default 'crisp' version, the algorithm is basically the same as the Generalised Minimum Volume Ellipsoid method described by Jolion et al in [169], with the difference that the MVE estimator is substituted by the MCD method, which has better statistical properties and is less computationally demanding than the former. The clustering algorithm iteratively partitions the space into clusters without a priori information about their number. At each iteration, the MCD estimator is run several times, corresponding to as many hypotheses about the fraction of points belonging to the cluster. The resulting cluster hypotheses are checked against 'known' distributions (typically multivariate gaussian distributions) by means of a Kolmogorov-Smirnov test [154]. The best fitting cluster is removed and a new iteration starts.

If the clusters are not well separated or 'crisp', cluster estimation can certainly be expected to degrade when there is a high number of overlapping clusters. All the points within a delineated cluster are removed from the feature space when in fact they could belong to the tail of an adjacent cluster. Along these lines, small changes in the shape of the large clusters extracted at the beginning of the algorithm can result in significantly different numbers of small clusters extracted at later iterations. Due to these factors, an alternative implementation of the algorithm has been developed where belonging to a cluster is interpreted in a fuzzy sense (e.g. [152, 170]) through the consideration of probabilistically-based membership functions.

A.2 Generalised minimum covariance determinant clustering

Let $X$ be a set of $n$ data points in an $m$-dimensional feature space, where every point has associated with it a weight $q_i$. The quantity $Q = \sum_{i=1}^{n} q_i$ can be regarded
as the "mass" of the feature space. If we have a set of data points $X_l$ at the $l$-th iteration, the best cluster is delineated and removed, yielding the new set $X_{l+1}$. The process stops when the number of remaining points becomes less than an assumed minimum cluster size, or when the number of detected clusters exceeds an upper bound. To extract a cluster, the space $X_l$ is analysed for different inclusion rates $h$, i.e. the assumed fractions of mass belonging to the cluster. The algorithm is termed generalised minimum covariance determinant (GMCD) clustering by analogy with the GMVE algorithm, since several values of $h$ are employed, whereas the original MCD (MVE) considers just one, typically $h = 0.5$, and $q_i = 1$.

The MCD algorithm finds the most compact cluster that contains the desired fraction $h$ of the mass of $X_l$ ($Q_l = \sum X_l q_i$) by minimising the determinant of the covariance matrix estimate over a number of random subsets of the original $m$-dimensional feature space (see chapter 5 for more details). Thus, we obtain estimates $T$ and $S$ (section 5.3) for the location and covariance matrix of the cluster respectively. We can define new estimates:

$$
T^0_h = T \\
C^0_h = \frac{d_{TS}}{\chi^2_{m,h}} S
$$

where the subscript $h$ indicates that the estimation was performed for an inclusion rate $h$, and the superscript 0 indicates that the quantities are only initial estimates. $C^0_h$ is a scaled version of $T$, which is weighted by the chi-square correction factor and by the biggest Mahalanobis distance corresponding to the fraction $h$ of the space mass for consistency with multivariate normal data (c.f. section 5.4).

Finally, if weights are defined for every data point $x_i$, $i = 1, \ldots, n$:

$$
\omega_i = \begin{cases} 
1 & \text{if } d_{TS}(x_i) = (x_i - T^0_h)^T(C^0_h)^{-1}(x_i - T^0_h) \leq \chi^2_{m,0.975} \\
0 & \text{otherwise}
\end{cases}
$$

the cluster parameters can then be recomputed using the weighted estimators

$$
T_h = \frac{\sum_{i=1}^{n} \omega_i q_i x_i}{\sum_{i=1}^{n} \omega_i q_i} \\
C_h = \frac{\sum_{i=1}^{n} \omega_i q_i (x_i - T_h)(x_i - T_h)^T}{\sum_{i=1}^{n} \omega_i q_i - 1}
$$

The estimators in Eq. A.3 are optimum in the least squares sense and thus improve the efficiency of the algorithm.

To determine the best cluster hypotheses among the estimates made for each value of $h$, the Kolmogorov-Smirnov test [154] is used. The Kolmogorov-Smirnov test compares the theoretical and the observed cumulative distribution functions of the Mahalanobis distances of the points contained inside the cluster using a quantity $D_h$, $D_h = max_i |F_{obs}(i) - F_{thr}(i)|$, as a test criterion.
The observed cumulative distribution function is defined as

\[
F_{obs}(i) = \begin{cases} 
0 & \text{if } i = 0 \\
\frac{\sum_{j=1}^{i} q_j}{\sum_{j=1}^{i_{\text{max}}} q_j} & \text{if } 1 \leq i \leq i_{\text{max}} \\
1 & \text{if } i > i_{\text{max}} 
\end{cases}
\]  
(A.4)

where \(i_{\text{max}}\) is the number of points contained inside the ellipsoid

\[
\Omega_{h} : (x_i - T_h)^T \Sigma^{-1}_h (x_i - T_h) < \chi^2_{m,0.975} 
\]  
(A.5)

Under the Gaussian cluster assumption, the theoretical cumulative distribution values \(F_{\text{thr}}(i)\) can be calculated from the relation

\[
d_{[T_h,C_h]}^{(i)} = \chi^2_{m,F_{\text{thr}}(i)} \quad i = 1, \ldots, i_{\text{max}} 
\]  
(A.6)

with \(d_{[T_h,C_h]}^{(i)}\) being the \(i\)-th lowest Mahalanobis distance in the cluster.

The significance level returned by the Kolmogorov-Smirnov test is used as the cluster quality measure and the best cluster hypothesis \(h = h^*\) is selected accordingly, thus obtaining the final cluster parameters: \(\mu_I = T_{h^*}, \Sigma_i = C_{h^*}\).

### A.3 Fuzzy version

When the clusters show a significant degree of overlap, it can be advantageous to introduce a reestimation step whereby the points allocated to each of the already extracted clusters are used to refine cluster estimates, instead of being removed forever.

Let us assume that \(l\) clusters had already been extracted and the corresponding points removed, and that a new cluster has been found at the \((l + 1)\)-th iteration.

Again, under the Gaussian cluster assumption, the following membership functions \(u_j\) (denoting membership to the \(j\)-th cluster) can be defined for each point \(x_i\) in the original set \(X\):

\[
u_j(x_i) = \begin{cases} 
\frac{v_j(x_i)}{\sum_j v_j(x_i)} & \text{if } v_j(x_i) > 0, \forall j \\
0 & \text{if } v_j(x_i) = 0, \forall j 
\end{cases}
\]  
(A.7)

with \(v_j(x_i)\) defined as follows:

\[
v_j(x_i) = \begin{cases} 
|\Sigma_j^{-1}|^{\frac{1}{2}} \exp\left(-\frac{1}{2} d_{[\mu_j,\Sigma_j]}(x_i)\right) & \text{if } d_{[\mu_j,\Sigma_j]}(x_i) < \chi^2_{m,0.975} \\
0 & \text{otherwise}
\end{cases}
\]  
(A.8)
Thereafter, the cluster parameters \( \mu_j \) and \( \Sigma_j \) \((j = 1, \ldots, l + 1)\) are reestimated; all points in \( X \) are reclassified and those whose Mahalanobis distances to any of the clusters lie within the 0.975 confidence level are removed (temporarily) before the following iteration.

### A.4 Experimental results and discussion

The two versions of the described algorithm have been used for colour model extraction in the XM2VTS [9] through chromaticity clustering (see chapter 4). Examples are shown for two clients: the first case (client 10, Figure A.1) corresponds to a situation where chromaticity cluster separability is fairly good, whilst in the second (client 25, Figure A.2) there exists considerable overlap between clusters. At each algorithm iteration, the following inclusion rates were considered: \( h = 0.5, 0.25, 0.125, 0.0625 \). The algorithm stops when only 5\% of the mass of the original feature space remains to be assigned to a cluster.

The default version, implicitly based on crisp clusters, usually yields 'cleaner' results when the clusters are relatively well separable, resulting in quite a distinct segmentation of the lip area from the surrounding skin area. The fuzzy version typically

---

**Figure A.1: Client 10, 1st session: good chromaticity cluster separability**

(a) Original image  
(b) Crisp clustering  
(c) Fuzzy clustering  
(d) Chromaticity histogram  
(e) Extracted clusters (crisp)
A.4. Experimental results and discussion

Figure A.2: Client 25, 1st session: poor chromaticity cluster separability
Appendix A. Robust clustering

performs better when there is significant overlap between the clusters and the variances are not too dissimilar between them. However, it may not yield such good looking when some clusters have significantly larger variances than others, since large variance clusters eventually 'absorb' a good number of points initially assigned to other clusters extracted earlier; likewise, the reestimation of the latter results in smaller variances, thus aggravating the problem.
Appendix B

Automatic lip tracking initialisation

B.1 Introduction

In order to both ease lip tracking performance from the very beginning and also reduce the influence of transients (by keeping to a minimum the number of frames required by the tracker to lock onto the lip contour) in a speaker verification framework based on lip tracking results, a relatively accurate initialisation is absolutely advantageous. For initialisation it is meant the process of locating the mouth area in an image and of producing a first B-spline approximation of the lip contour.

Tracking initialisation was, however, left aside, shadowed by the developments in tracking proper. Despite the promising results reported in [138] the fact is that human-aided initialisation (by manually cropping a rectangular window containing the mouth area, as described in chapter 4 for the purpose of extracting colour models) has been common practice. Nevertheless, the realisation of the tracking experiments on the XM2VTS database [9] according to the Lausanne protocol (chapter 7), which required tracking a few thousands of sequences, and therefore the same number of initialisations, motivated a renewed interest in automating initialisation as much as possible and, consequently, in the revision of the method described in [138] in a new context that further extends the applicability of the developments in robust estimation and clustering that have already been described in chapter 5 and in Appendix A.

The method consists of three steps. First, colour information is exploited for locating the face and approximating its shape by an ellipse. In a second step, intensity gradient projections are used to determine a number of lip candidate positions since lips are characterised by a strong projection value of the derivatives along the normal to their boundaries. This criterion, however, is not enough to determine the position of the lips since other face regions such as the eyebrows or the skin area surrounding the nostrils, for instance, exhibit the same property, apart from the presence in some other images of glasses, moustaches, and beards. As a result, a final process is
Appendix B. Automatic lip tracking initialisation

Figure B.1: Face location and approximation by elliptical shape

aimed at evaluating the goodness of the several candidates found and deciding which candidate is most likely to correspond to the mouth. In this third step, the candidate hypotheses are tested using colour models and ranked accordingly to select the best one(s).

B.2 Face location and approximation by an elliptical shape

Assuming that human faces can be reasonably well modelled by oval shapes, face location is posed as the problem of finding the ellipse that best fits the spatial distribution of pixels that correspond to a ‘skin’ chromaticity model.

In the current implementation, the problem is solved by seeking the best elliptical cluster (see Appendix A) that contains at least 50\% (h=0.5) of the total “mass” $Q$ of the image, with $Q = \sum q_i$, and $q_i$ is the probability that pixel $(x_i, y_i)$ belongs to the ‘skin’ chromaticity class corresponding to a given subject (obtained by averaging the models computed for that subject across sessions). For the Kolmogorov-Smirnov test, the theoretical probability distribution function is supposed to be a uniform, elliptical distribution.

The process is illustrated in Figure B.1. The original image is shown in Figure B.1(a). Figure B.1(b) shows those pixels whose Mahalanobis distance to the skin colour model was within a 0.975 confidence level together with the resulting best fitting ellipse.

In an alternative implementation, a generic skin colour model $(\mu_G, \Sigma_G)$ (obtained by averaging the skin colour models of a random sample of 100 images of different subjects from the XM2VTS database [9]) was used as a seed colour model. The robust clustering algorithm of Appendix A decomposes the original image into a number of chromaticity clusters, and the most similar $(\mu_s, \Sigma_s)$ to the generic model
B.3. Generation of lip candidates based on grey-level gradient projection

(i.e. the one with the lowest Bhattacharyya [145] distance) is selected and eventually combined with the generic one in order to generate an adapted model \((\mu_A, \Sigma_A) \) \((\mu_A = \frac{\mu_G + \mu_S}{2}, \Sigma_A = \Sigma_G)\), that is, the mean vector is translated, whilst the covariance matrix stays the same to avoid gross variance terms) which is used instead of the above subject-specific model. Not surprisingly (recall the considerations made in chapter 4 about skin colour), better results were obtained with subject-specific models, which can be computed beforehand and therefore be advantageous in that colour clustering can be spared.

B.3 Generation of lip candidates based on grey-level gradient projection

Once we have obtained a region corresponding to the face area, the subsequent process is restricted to such a region looking for lip-like regions inside it. The process followed takes into account the saliency of the gradient projection along the normal direction to the lip boundaries. Similar approaches based on vertical projection of horizontal edges are described, for instance, in [6, 49].

The original RGB image is converted into an intensity image and its gradient is computed using Sobel masks [171], so that some smoothing is obtained in the perpendicular direction to the one the gradient is computed along.

\[
G_x(i,j) = I_{i,j+1} + 2I_{i,j} + I_{i,j-1} - [I_{i-1,j-1} + 2I_{i-1,j} + I_{i-1,j+1}] \quad (B.1)
\]

\[
G_y(i,j) = I_{i+1,j-1} + 2I_{i+1,j} + I_{i+1,j+1} - [I_{i-1,j-1} + 2I_{i-1,j} + I_{i-1,j+1}] \quad (B.2)
\]

\(G_x\) and \(G_y\) are respectively the derivative masks along the \(x\) and \(y\) directions and \(I_{k,l}\) is the intensity value of the pixel \((k,l)\).

Gradient images are then obtained by replacing each pixel with the magnitude of its directional intensity derivative along the \(\vec{l}\) direction, i.e.

\[
D(i,j) = |\nabla I_{i,j} \cdot \vec{l}| \quad (B.3)
\]

The magnitude values of the computed directional derivative are then projected onto the ellipse main axis:

\[
P(x, y) = \sum_k D((x, y) + k\vec{n}) \quad (B.4)
\]

\(\vec{n}\) being a unit vector orthogonal to \(\vec{l}\). In practice, \(\vec{l}\) corresponds to the vertical direction (in-plane face rotations can be considered negligible).

In Figure B.2(a) we can see the gradient projection profile for the previous image. The lips will lie around a local maximum of the gradient projection occurring at a
Appendix B. Automatic lip tracking initialisation

Figure B.2: Generation of lips location hypotheses using gray-level gradient projections

Position on the main axis of the ellipse approximating the face shape. The mean gradient projection value is computed and all the local maxima above it are analysed to determine which one correspond to the lip position.

Gradient projection maxima are relocated to correspond to positions along the main axis of the best fitting ellipse. Around these centered positions, boxes are automatically created, the dimensions of which are related to the estimated face size. The resulting boxes for the present case are shown overlaid in Figure B.2(b).

B.4 Selecting the best candidate

Analogously to face location, the selection of the best candidate hypothesis is based on two aspects: a) colour model compliance, and b) spatial distribution. In each rectangular box, the best ellipse fitting the spatial distribution of lips-like pixels is sought after and a fitness measure is defined by the quotient between the "mass" of pixels inside the ellipse and its area, that is, the mass density. This measure, which
B.5 Discussion

The method described above relies on very simple assumptions and should be understood as a procedure for helping tracking automation under human supervision that extends further the applicability of the robust estimation techniques described in chapter 5 and in the Appendix A.

Whereas the ellipse hypothesis works reasonably well to model the spatial distribution of skin-like pixels in the facial area, this same simple hypothesis proves a bit weak for modelling the lips area. The colour of the lips is not always very distinctive.
from the surrounding skin area, and in fact some other areas in the face happen to exhibit similar colour characteristics. This is why a shape checking stage is enforced. However, the fact the lips area is only very coarsely approximated by an elliptical distribution results in that some occasional blobs with lip-like chromaticities can sometimes provide a better hypothesis than the proper lips.
Appendix C

Robust facial characterisation and reconstruction

C.1 Introduction

Appearance-based methods based on principal component analysis (PCA), also known as Karhunen-Loève transform (KLT), are quite common in the face recognition literature as a means of representing faces in a compact way by exploiting their statistical variability, as originally described in [38], where eigenpictures are generated to obtain a low-dimensional, yet accurate enough, representation of human faces.

In the most common scenario, faces are modelled as a linear combination of intensity modes of variation, the so-called eigenfaces [36], learnt from a large, representative, database. The Karhunen-Loève transform can be understood as a first order approximation of the statistical distribution of faces. For instance, in [41] the space of faces is characterised as a multivariate gaussian probability density function used for the detection of face-like objects. This compact representation has also been used for recognition, where face matching is done by comparison of the coefficients of the KLT expansion using some suitable metric [172, 39, 66, 173].

Under general conditions, it seems that facial appearance would lie in a continuous manifold where, as pointed out in [72, 21], the variability for a given face across changes in illumination, viewpoint and expression is greater than it is between different faces when these three factors are held constant. As a result, considerable research efforts have been focused in analysing and compensating for these factors, especially lighting [174, 175], for which it is claimed [72] that images of a particular face under varying illumination would lie in a 3-D linear subspace of the high dimensional feature space.

If we concentrate on the case of frontal faces obtained under specified lighting conditions, and overlook the fact that a simple superposition of grey-scale modes of variation misses out face shape variability associated with individual appearance
and expression changes, then the expansion in eigenfaces can still be a valid approximation of the space of faces.

The coefficients of the expansion are usually obtained through projection of the target face onto each of the existing eigenfaces. This method, however, is not robust, as it cannot handle problems related to occlusion or, more generally, to the presence of outliers of the face space.

As a first step, the problem of reconstructing face images from incomplete measurements is analysed. Sampling techniques have been successfully reported for face localisation [56] as a means of reducing the computational cost of the process. In this case, however, the motivation is the robust characterisation of objects from incomplete (and possibly contaminated) data.

When outliers can be identified and discarded beforehand, it is shown how the estimation of the coefficients of the expansion in eigenfaces boils down to solving a simple algebraic problem that admits a trivial closed-form solution.

The main problem is, however, how to automatically detect those outlier measurements on the fly and the attention was drawn to the statistical characterisation of the space of faces in order to evaluate pixel discrepancies. Based on the inspiring work [176] by Leonardis and Bischoff, a novel method is proposed that addresses coefficient estimation as a robust optimisation problem. In this novel method an analysis of both pixel error distributions and a pixelwise error model is exploited to conduct the optimisation routine aimed at estimating the coefficients.

C.2 Face space modelling

In our current eigenface framework, we deal with faces automatically registered (for instance, using the maximum-likelihood face detection method described in [41]). Following geometrical and photometrical normalisation, and background suppression, a face $\mathbf{x} = [x_1, \ldots, x_N]^T \in \mathbb{R}^N$ (where the $x_i$ are the grey-scale values of the normalised $l$-by-$m$ input image arranged in raster order and $N = l \times m$ is the number of pixels of the original image), can be approximated by a linear combination $\mathbf{\hat{x}}$ of eigenfaces $\mathbf{e}_j$ (we will assume all faces have had the average face subtracted from them):

$$\mathbf{x} \simeq \mathbf{\hat{x}} = \sum_{j=1}^{n} a_j(\mathbf{x}) \mathbf{e}_j$$  \hspace{1cm} (C.1)

where the eigenfaces $\mathbf{e}_j$ are the eigenvectors of a maximum-likelihood estimate of the face covariance matrix, and $n$ is the dimensionality of the face space ($n < N$), i.e. the number of eigenvectors retained to represent faces to a sufficient degree of accuracy.

The above linear representation is optimal in the sense of minimising the reconstruction error $\|\mathbf{x} - \mathbf{\hat{x}}\|^2$. Typically the coefficients $a_j$ of the expansion in eigenfaces are computed through projection of the target object onto each of the $\mathbf{e}_j$, i.e.
C.3. Reconstruction from incomplete measurements

C.3.1 Sample-based least squares estimation

The best approximation of $\mathbf{x}$ in the $n$-dimensional eigenspace is given by:

$$ \mathbf{x} \simeq \sum_{j=1}^{n} a_{j}(\mathbf{x}) \mathbf{e}_{j} \quad (C.4) $$

In particular, the $k$-th pixel of $\mathbf{x}$, $x_k$, would be approximated by:

$$ a_j(\mathbf{x}) = \langle \mathbf{x}, \mathbf{e}_j \rangle = \sum_{i=1}^{N} x_i e_{ji} \quad (C.2) $$

or, in a more compact matrix notation:

$$ \mathbf{a}_0 = \mathbf{E}^{T} \mathbf{x} \quad (C.3) $$

where $\mathbf{a}_0$ represents the vector of eigenface coefficients, and $\mathbf{E}$ is an $N \times n$ matrix consisting of the first $n$ eigenfaces arranged in columns.

As reported in [176], this form of computing the coefficients is quite sensitive to spurious errors and, in particular, to occlusion. Although errors are filtered to some extent by the KLT expansion, the errors are not localised at the occluded part and rather spread across the whole vector of $a_j$, as can be seen in Fig C.1. The first 200 coefficients of the eigenspace expansion were used for reconstructing the face in this example. The eigenfaces were computed using a large training set from the FERET database [75].

Figure C.1: Effect of occlusion in image reconstruction
Appendix C. Robust facial characterisation and reconstruction

\[ x_k \simeq \sum_{j=1}^{n} a_j(x) \epsilon_{jk} \]  
(C.5)

So, if we consider a sample of \( r \) pixels from the original image, \( (x_{i_1}, \ldots, x_{i_r}) \), we will have:

\[
\begin{pmatrix}
  x_{i_1} \\
  \vdots \\
  x_{i_r}
\end{pmatrix}
= \begin{pmatrix}
  \epsilon_{1i_1} & \cdots & \epsilon_{ni_1} \\
  \vdots & \ddots & \vdots \\
  \epsilon_{1i_r} & \cdots & \epsilon_{ni_r}
\end{pmatrix}
\begin{pmatrix}
  a_1 \\
  \vdots \\
  a_n
\end{pmatrix}
\]  
(C.6)

or, in a more compact matrix notation:

\[ x_r \simeq E_r a \]  
(C.7)

Finding \( a \) can be posed as a least squares problem \(^1\), where we look for the coefficients which minimise the reconstruction error over the pixel sample set:

\[ \epsilon_r = (x_r - E_r a)^T (x_r - E_r a) \]  
(C.8)

for which the solution is \( a = E_r^+ x_r \), where \( E_r^+ = (E_r^T E_r)^{-1} E_r^T \), which boils down to Equation C.3 when all the samples are used in the computation.

C.3.2 Estimation error and reconstruction quality

A measure of the dependence of the estimation error on the number of measurements used can be obtained by computing the Mahalanobis distance \( \xi \) between the coefficient vector \( a \) estimated as described above for a given sample size, and \( a_0 \), obtained through direct projection onto the eigenspace, i.e.,

\[ J = (a - a_0)^T \Lambda^{-1} (a - a_0), \]

where \( \Lambda \) is a diagonal matrix consisting of the first \( n \) eigenvalues.

If we recall the expressions to compute both \( a_0 \) and \( a \):

\[ a_0 = E^T x \]  
(C.9)

\(^1\)The coefficient vector can also be estimated within a maximum a posteriori (MAP) framework, by finding the \( a \) that maximises the a posteriori probability \( P(a|x_r) \). If we consider:

\[ x_r = E_r a + v_r \]

where \( v \) is the sample reconstruction error vector, which we shall model as an additive, isotropic noise term with a multivariate normal probability density function, \( v \equiv N(0, \sigma^2I) \). Likewise, assuming that the probability density function of \( a \) is also a multivariate Gaussian function, \( a \equiv N(0, \Lambda) \), it is easy to show that the optimal \( a \) would be given by:

\[ a = (E_r^T E_r + \sigma^2 \Lambda^{-1})^{-1} E_r^{-1} x_r \]

Obviously, in the absence of noise the MAP approach would produce identical solutions to the obtained by the least squares method.
C.3. Reconstruction from incomplete measurements

Figure C.2: Coefficient estimation error as a function of the sample size

\[ a = (E_r^T E_r)^{-1} E_r^T x \]  \hspace{1cm} (C.10)

and consider an \( r \times N \) selecting matrix \( G \), such that there is only a unit-valued element in each row, in a column corresponding to a selected measurement of \( x \), all the other elements being zero, we can write:

\[ x_r = Gx \]  \hspace{1cm} (C.11)

and analogously for the eigenfaces

\[ E_r = GE \]  \hspace{1cm} (C.12)

so that Equation C.10 can be rewritten as

\[ a = (E^T (G^T G) E)^{-1} E^T (G^T G)x \]  \hspace{1cm} (C.13)

Taking into account the definitions above, \( J \) can be rewritten as:

\[ J = x^T M x \]  \hspace{1cm} (C.14)

where \( M = [(G^T G) E (E^T (G^T G) E)^{-1} - E] \Lambda^{-1} [((E^T (G^T G) E)^{-1} E^T (G^T G) - E^T]. \)

The expected value of \( J \), denoted as \( \bar{J} \), is given by

\[ \bar{J} = Tr[\Sigma M] \]  \hspace{1cm} (C.15)

with \( \Sigma \) being the \( N \times N \) covariance matrix of the input images.
Appendix C. Robust facial characterisation and reconstruction

Figure C.3: Examples of image reconstruction for different sample sizes

If $\Sigma$ is no longer available following eigenface computation, an estimate $\hat{\Sigma}$ can be computed from $\Lambda$:

$$\hat{\Sigma} = (\mathbf{EE}^T)^+\mathbf{EAE}^T = \lim_{\epsilon \to 0} (\mathbf{EE}^T + \epsilon \mathbf{I})^{-1}\mathbf{EAE}^T$$  \hfill (C.16)

$J$ reflects estimation discrepancies as a consequence of sampling, as can be seen from the dependence on the selecting matrix $\mathbf{G}$. In the limit, when all pixels are included in the sample, the estimation error drops to zero.

Figure C.2 shows the effect of increasing the sampling density on the average estimation error$^2$ ($\pm 1\sigma$) for one of the subjects in the database. As can be expected, the estimation error diminishes steadily with the increase of the sampling density, and is practically negligible even if just one third of the pixels are used for the estimation, as can be seen in Figure C.2 and also in Figure C.3, where some image reconstruction examples are shown for different sample sizes.

C.3.3 Introducing occlusion

Figure C.4 illustrates the effect of outliers on coefficient estimation in a simple 2D space where the data can be approximated by a 1D linear subspace. Point P would correspond to a typical object for which the best approximation in the subspace would be given by its orthogonal projection $Q$. Nevertheless, if the ordinate measurement $y$ is extraordinarily poor, as shown by point P1 in the figure, the projection onto the subspace, given by $Q1$, would result in a substantial estimation error. On the other hand, if we identify the ordinate measurement as wrong and decide to substitute its value by its mathematical expectation ($y = 0$), we would be at point P2, for which we would obtain point Q2. Obviously, even if we have identified the ordinate measurement as an outlier, we are not making the best use of the information.

$^2$20 tests were run for each sample size considered: 300, 500, 1000, 2000 and 3000 pixels. Twenty trials represent a very sparse sampling of matrix $\mathbf{G}$ for a given sampling cardinality, but still enough to make an impression. It is worth noting that, although the test images are $128 \times 128$ pixels, the actual face area amounts to slightly more than 3000 pixels.
C.4 Coefficient estimation as a robust optimisation problem

Figure C.4: Estimating the eigenspace coefficients

available to find the best object approximation in the subspace. Taking into account the underlying data structure the best projection onto the subspace would be given by $R$.

The approach described in the previous section has been applied to the problem of reconstruction of occluded images. A synthetic occluding pattern like the one shown in Figure C.1 was superimposed onto the test images, removing nearly 25% of the total facial area. Coefficient estimation was performed by using the remaining pixels as valid measurements. Typical results for some of the approximately 100 images tested can be seen in Figure C.5. The first column shows original, unoccluded images, and the second one their projections onto the eigenspace. The effect of applying the same principle to occluded images can be seen in the third column. The fourth column shows the effect of substituting outlier pixels by their values in the average face prior to projection. Finally, the fifth column shows image reconstruction using the described least squares estimation. As can be seen, the results show significant improvements over the reconstructions based on direct projection of the input image onto the eigenspace, and provide quite natural-looking reconstructions for the tested images.

C.4 Coefficient estimation as a robust optimisation problem

As described in [176], the coefficients of the eigenface expansion can be computed by sampling a high enough number of pixels of the input image and minimising the pixelwise reconstruction error over the subset of sampled points. Coefficient estimation is posed as the solution, in the least squares sense, of the following error function:
Figure C.5: Examples of image reconstruction: P) original, unoccluded image; Q) projection onto the eigenspace; Q1) occluded, projection; Q2) occluded, projection after outlier substitution; R) occluded, least squares estimation.
C.4. Coefficient estimation as a robust optimisation problem

\[ \varepsilon^2(r) = \sum_{i=1}^{k} (x_{r_i} - \sum_{j=1}^{n} a_j(x)e_{j,r_i})^2 \]  \hspace{1cm} \text{(C.17)}

where \( r = (r_1, \ldots, r_k) \) is a set of image points obtained by random sampling of the input image.

The minimisation of Eq C.17 will only produce correct \( a_j \) values if the set of points \( r_i \) does not contain outliers, i.e. noisy pixels or points belonging to occluded areas. As a result, \( \varepsilon(r) \) has to be minimised in a robust manner whereby we take into account the pixelwise error distribution to prune out outliers.

In the inspiring work by Leonardis et al [176], it is reported that would-be outliers are pruned out from the sample set, but not many details are provided about how this is actually done with regard to the error distribution, and how the ultimate set over which the the reconstruction error is minimised is generated. As we shall see this aspect has its implications in the sense of modifying the resulting error distribution, with clearly noticeable effects in the reconstruction of human faces that maybe are not so important with the kind of objects being dealt with in [176].

Computing the coefficients of the expansion in eigenfaces can be posed as a robust regression problem [177, 178]. Another formal attempt at solving robust regression problems like this can be found in [155]. The robust estimator suggested therein is based on LMedS, i.e. least median of square residuals. Notwithstanding the high breakdown point [151] of the estimator, its efficiency can be relatively low in the presence of additive gaussian noise, which is quite the case of the reconstruction residuals of the expansion in eigenfaces, as will be shown below. In connection with this, we shall also see that ‘conservative’ guesses about the degree of occlusion -i.e. basing the estimation on only the ‘best’ half of the data- results in poorer estimates than when as many ‘good’ points as possible can be used. A final reservation about the method suggested in [155] has to do with the huge number of sampling sets required, as it happens when the RANSAC method [153] is applied to high dimensionality problems.

In the RANSAC paradigm, the parameters are estimated from a minimum number of required data. Instead of using as much data as possible, RANSAC generates a number of ‘small’ random sets, wherefrom the characteristic parameters are computed and data consensus analysed. Unfortunately, in our ‘high’ dimensional space, the computational cost is far more than acceptable. The expected number of random sets to be generated is proportional to \( w^{-n} \), where \( w \) is the probability of drawing a valid data point and \( n \) is the cardinality of each random set. For instance, if every set consists of 200 points drawn at random, and 10\% of the image is occluded, the figure is of the order of \( 10^9 \).

C.4.1 Bringing pixel error distributions and error expectations together

If the fraction of non-outlier pixels in the image is \( \alpha \), and we know the ‘real’ eigenface coefficients representing the original image, we can expect smaller reconstruction
errors for that fraction $\alpha$ of pixels. Taking into account the considerations about estimation errors made in section C.3, we can also expect that, if there is another set of coefficients that better represent that fraction $\alpha$ of non-outlier pixels, the estimation error can be kept without reasonable bounds, the more so the higher $\alpha$ (less occlusion) and the lower the measurement noise (the data distribution is well modelled by the eigenfaces).

Accordingly, the criterion function eventually used has been defined as the sum of the best $\alpha N$ pixel reconstruction errors.

$$J = \sum_{i=1}^{\alpha N} \epsilon_i^2$$ (C.18)

where $\epsilon_i$ is the $i$-th biggest pixel reconstruction error in magnitude.

The results of the optimisation algorithm (once again, Dynamic Hill Climbing was used [156, 157]) will then depend on the prior knowledge available about the percentage of outliers in the test image. In Figure C.6 we can see the reconstructions obtained (200 eigenfaces) for both the original and the occluded image, for different assumptions made about the degree of occlusion. The best results are attained when there is a good match between the occlusion hypothesis, i.e. when $\alpha = 1.0$ for the original image, and when $\alpha = 0.75$ in the case of the occluded image, which is about right.

As we can notice, wrong occlusion hypotheses result in wrong coefficient estimation. When the degree of occlusion is underestimated, outliers will still be expected to play a role in estimation and distort the computation. In the opposite case, when the occlusion is overestimated the solutions tend to overfit the reconstruction of the best fraction $\alpha$ of pixels at the expense of failing to approximate the rest of 'good' pixels.

At this stage we have reduced the level of prior knowledge required to reconstruct the images (we do not have to manually identify and remove the occluding area) but we still require to have a good guess about the fraction of outliers in the image. However, it is possible to run the optimisation routine for several occlusion hypotheses, and then check its normalised histogram of reconstruction errors.

It is possible to model the error distribution as a zero-mean gaussian function of standard deviation $\sigma$, where, in principle, $\sigma$ should take some value close to $\lambda_{n+1}^{\frac{1}{2}}$ [179] (in our case $\lambda_{n+1}^{\frac{1}{2}} \sim 20$). Alternatively, an expectation of the real distribution could be learnt from a set of images. The derived error expectation actually overestimates the variance of the distribution, but it can be used as a good threshold beyond which bigger errors are extremely unlikely.

For each occlusion hypothesis, we would get an 'optimal' solution for which we will obtain a certain error distribution. This error distribution can then be converted into a normalised histogram and matched against the theoretical error distribution, i.e. we could consider a distortion measurement $\xi$.  


C.4. Coefficient estimation as a robust optimisation problem

Figure C.6: Image reconstructions for different hypotheses about the degree of occlusion in the test image
Appendix C. Robust facial characterisation and reconstruction

\[ \xi = \int_0^{\lambda_{h+1}} (f_{th}(\epsilon) - f_\alpha^*(\epsilon))^2 d\epsilon \]  
\[ f\_th \text{ is the theoretical error distribution (histogram) and } f\_\alpha^* \text{ is the normalised histogram for a given hypothesis } \alpha, \]

\[ f_\alpha^*(\epsilon) = \frac{1}{\int_0^{\lambda_{h+1}} f_\alpha(\tau) d\tau} f_\alpha(\epsilon) \]  
Figure C.7 illustrates the effect of the \( \alpha \) hypotheses on the error distributions. The error distribution in the absence of occlusion has been selected in this case to represent the 'theoretical' distribution. Although no numerical distortion values are presented it is clear from the figures how a mismatch between the occlusion hypothesis and the real percentage of outliers results in artifacts and distortions with reference to the expected distribution.

C.4.2 Occlusion hypothesis testing by matching the distribution of reconstruction residuals

As discussed in the preceding subsection, we can check the goodness of the coefficient estimates by matching the resulting pixelwise residual distribution with the characteristic distribution of the residuals in the absence of occlusion. In that case we used the residual distribution of the same image being tested, which is neither realistic nor general, as this can be somewhat dependent on the characteristics of the image, as shown in Figure C.8, where residual distributions are shown for a number of images. Hence, for the sake of generality, a residual distribution for testing was obtained from a pool of sample images.

The residual distributions of the approximately 70 images used for obtaining the test distribution look very similar, except for a few of them which exhibit a larger variance (see the two broadest distributions in Figure C.8 above) as a result of having some features like facial hair that were not present in the training database used to generate the eigenfaces. As a result, it was decided to run a robust estimator of location and scale on the pixelwise residuals of the approximately 70 images. The normalised residual histogram, together with gaussian fits corresponding to the different tried estimators, is shown in Figure C.9. The robust tanh location and scale estimator and the MAD (median absolute deviation) provided almost identical results, as opposed to the typical estimation of mean and variance of the data (mean-based estimation in the figure).

For occlusion hypothesis testing we have defined two criterion functions. The first one is the discrete version of Equation C.19:

\[ J_1 = \sum_{i=1}^{k^*} (f_\alpha^*(\epsilon_i) - f_{th}(\epsilon_i))^2 \]
C.4. Coefficient estimation as a robust optimisation problem

Figure C.7: Normalised error histograms for different occlusion hypotheses (top row: original image; bottom row: occluded image)
Figure C.8: Pixelwise residual distributions for a number of sample images

Figure C.9: Residual distributions and gaussian fit resulting of robust estimation
whilst a second criterion function is the adaptation of a chi-square test [154]:

\[
J_2 = \sum_{i=1}^{k^*} \frac{(f_o^*(\epsilon_i) - f_{th}(\epsilon_i))^2}{f_{th}(\epsilon_i)}
\]  

(C.22)

where \( k^* \) is such that \( \epsilon_i \leq \sqrt{\lambda_{n+1}} \).

Table C.4.2 shows the results of the tests for both the occluded and the unoccluded image, and for both criterion functions.

<table>
<thead>
<tr>
<th>occlusion hypothesis</th>
<th>unoccluded</th>
<th>image</th>
<th>occluded</th>
<th>image</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>( J_1 = 1.59 \times 10^{-2} )</td>
<td>( J_2 = 0.17 )</td>
<td>( J_1 = 3.83 \times 10^{-2} )</td>
<td>( J_2 = 70.30 )</td>
</tr>
<tr>
<td>25 %</td>
<td>( J_1 = 4.36 \times 10^{-2} )</td>
<td>( J_2 = 4.09 )</td>
<td>( J_1 = 2.26 \times 10^{-2} )</td>
<td>( J_2 = 0.92 )</td>
</tr>
<tr>
<td>50 %</td>
<td>( J_1 = 4.91 \times 10^{-2} )</td>
<td>( J_2 = 18.99 )</td>
<td>( J_1 = 5.04 \times 10^{-2} )</td>
<td>( J_2 = 14.00 )</td>
</tr>
<tr>
<td>75 %</td>
<td>( J_1 = 5.63 \times 10^{-2} )</td>
<td>( J_2 = 52.64 )</td>
<td>( J_1 = 5.63 \times 10^{-2} )</td>
<td>( J_2 = 52.64 )</td>
</tr>
</tbody>
</table>

It is possible to see how -specially with the chi-square test- the lowest values of the criterion function correspond to the best occlusion hypotheses (0% for the unoccluded image, and about 25% for the occluded one).

C.5 Discussion

It is possible to establish a parallelism between the presented hypothesis testing stage and the goodness-of-fit tests performed for robust clustering [152] (see Appendix A). The fact that we have a gaussian error model in this case should not be surprising, in the light of the central limit theorem [154].

The current limitations about image registration can be avoided: the optimisation routine would then search for affine-warped sets of coefficients that result in minimal (robust) reconstruction error. This could lead in fact to a routine for robust face detection and characterisation, which could then be subsequently used for matching or image reconstruction.

Although described in the framework of face recognition, the presented method is quite general and can be applied to any linear regression problems for which a statistical error model exists.

The results obtained solving this high-dimensional optimisation problem look very promising but the method should be tested more extensively and consider different forms of occlusion and image corruption. The narrower the statistical error distribution, and the lower the degree of corruption, the higher the prospects of success.
Bibliography


