A Thesis submitted for the degree of Doctor of Philosophy

Advisor: Dr. Peter Corcoran

College of Engineering & Informatics
National University of Ireland, Galway
May 2009

Advances in the Modeling of Facial Sub-Regions and Facial Expressions using Active Appearance Techniques

By: Ioana Bacivarov

Table of contents

List of Figures

List of Tables

List of Abbreviations

1A1	One-Against-One
1AA	One-Against-All
3DMM	3-D Morphable Models
AAM	Active Appearance Model
AAU	Anthropometric Action Unit
AMB	Anthropometric-Muscle-Based
ANN	Artificial Neural Networks
ASM	Active Shape Model
AU	Action Unit
BAM	Boosted Appearance Model
CCA	Canonical Correlation Analysis
COG	Centre of Gravity
DAG	Directed Acyclic Graph
DAM	Direct Appearance Model
DCM	Discrete Choice Models
EEG	Electroencephalograms
EFM	Enhanced Fisher Classification Models
EOG	Electrooculographic
FACS	Facial Action Coding System
FAME	Flexible Appearance Modelling environment
FLD	Fisher Linear Discriminant
FN	False Negatives
FP	False Positives
GDA	General Discriminant Analysis
HMM	Hidden Markov Models
HSV	Hue, Saturation, and Value
HVS	Human Visual System
ICA	Independent Component Analysis
IR	Infra-Red
LDA	Linear Discriminant Analysis

LUT	Look-Up-Table
LVQ	Learning Vector Quantisation
MLP	Multilayer Perceptron
MSA	Multi-Scale Analysis
NN	Nearest Neighbour
PAAM	Pairwise Active Appearance Model
PCA	Principal Components Analysis
PDM	Point Distribution Model
Pt-Crv	Point-to-Curve
Pt-Pt	Point-to-Point
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SVM	Support Vector Machines
TP	True Positives
TPS	Thin Plate Splines
TN	True Negatives

Acknowledgements

I would like to thank everybody that helped me during my Ph.D study at the National University of Ireland, Galway. First of all, I would like to express my sincere gratitude to my advisor, Dr. Peter Corcoran, for his guidance and support in academic research. Dr. Corcoran allowed me to benefit from his knowledge, experience, financial support, and many of his original views, while not constraining my creative efforts. He also patiently commented and reviewed this manuscript.

I am grateful to my PhD committee for their valuable ideas, suggestions, and encouragement.

This work was co-sponsored by FotoNation (Ireland) Ltd and Enterprise Ireland under the Innovation Partnership Program grant No. IP/2006/361. A special thanks is given to all my colleagues in FotoNation, in particular to my research manager Alexandru Drambarean, not only for their useful technical support, but also for the pleasant and friendly environment that they have offered me. A special mention to Anne, Claudia, Gabi, Norah (also for proofreading my thesis), Sharon, and Istvan. They offered me not only their support, but most importantly, their friendship. Our fun activities outside the office space made my stay in Ireland really enjoyable, indirectly inspiring my work.

I would also like to thank Mircea Ionita, a Ph.D fellow student, who conducted his thesis in parallel. His initial help, by introducing me into the AAM field and its software tools, gave me a really good start in my work. He also constantly gave me constructive feedback and had important contributions in many of our common published papers.

Last, but not least, I would like to express my deepest gratitude to my family. As both my parents and my sister, Iuliana, wrote Ph.D thesis, it allowed me to learn from their experiences. They genuinely encouraged me not only during my Ph.D, but during all my life experiences. Thank you for your unconditional love and support!

Finally, many thanks to all who over the last eight years contributed introducing me to the fascinating world of engineering by forming me as a researcher at the universities of Bucharest, Grenoble, and Galway

Thank you all!

Abstract

Advances are presented in the modelling of facial sub-regions, in the use of enhanced whole face-models, and based on these sub-region models, in the determination of facial expressions. Models are derived using techniques from the field of active appearance modelling (AAM). A technical description and review of such techniques and a number of additional state-of-art techniques for face detection and face region analysis are provided. A detailed literature review covering a range of topics relating to facial expression analysis is provided. In particular the prior use of AAM techniques for facial feature extraction is reviewed. A range of methodologies for classifying facial expressions are also reviewed. Improved eye-region and lips-region models are presented. These models employ the concept of overlapping landmark points, enabling the resulting models to handle eye-gaze, different degrees of closure of the eye, and texture variations in the lips due to the appearance of teeth when the mouth opens in a smile. The eye model is further improved by providing a component-AAM implementation enabling independent modelling of the state of each eye. Initialisation of the lips model is improved using a hue-based pre-filter. A whole-face component-AAM model is provided, combining the improved eye and lips models in an overall framework which significantly increases the accuracy of fitting of the AAM to facial expressions. A range of experiments are performed to tune and test this model for the purpose of the accurate classification of facial expressions in unseen images. Both nearest neighbour (NN) and support vector machine (SVM) classification methodologies are used. Testing of the system to classify the six universal emotions and the neutral face state shows that an accuracy of 83% can be achieved when using SVM classification. Preliminary investigations on additional enhancements to improve on this performance are provided, including the use of (i) pre-filters for gender, race, and age, (ii) person-specific AAM models, and (iii) expression tracking across multiple images in a video sequence. All of these techniques are shown to have potential to further enhance the accuracy of expression recognition of the underlying component-AAM face model with eye and lips sub-regional models.

Chapter 1 - Introduction

1.1. Background and Motivations

In the recent past, a digital camera was only able to capture images. Recently, however, face-tracking technology has become a common feature of consumer digital cameras. The most recent implementations feature hardware coded face tracking in an IP-core [1]. Applications to date have been limited to optimising the exposure and acquisition parameters of a final image [2]. Yet there are many additional applications which have even greater potential to enrich the user experience.

The rapid deployment of face tracking technology in cameras suggests that other more advanced face analysis techniques will soon become feasible and begin to offer even more sophisticated capabilities in such consumer devices.

The detailed analysis of facial expression is one such technique which can offer a wide range of new consumer applications for mobile and embedded devices. In the context of managing our personal image collections it is useful to be able to sort images according to the people in those images [3]. It would be even more useful if it were possible to determine their emotions and thus enable images to be further sorted and categorised according to the emotions of the subjects in an image.

Other consumer device applications also can gain from such capabilities. Many such devices now feature a camera facing the user, e.g. most mobile smart-phones, and thus user-interfaces could respond directly to our facial expressions. An e-learning system could match its level of difficulty to the degree of puzzlement on the student's face. A home health system could monitor the level of pain from an elderly person's facial expression [4, 5]. Other domains such as entertainment and computer gaming, automotive [6], or security [7] can also benefit from such applications. As the underlying expression recognition technologies improve in accuracy, the range of applications will grow further.

The research undertaken in this thesis represents strategic medium-term research for the sponsoring company, FotoNation (Ireland) Ltd. This was originally focussed on face analysis and recognition techniques, and in particular on AAM facial models which represent practical technology that the company considered as having potential for commercialisation as a second-stage development of its in-camera face

tracking technology. As the research progressed it was realised that more sophisticated applications could be realised using AAM modelling techniques.

Thus the core research theme of this doctoral thesis began to emerge. After our initial successes with improving the underlying AAM face model it was decided to pursue studies of some more ambitious and sophisticated models to determine the practicalities of modelling of unconstrained faces with sufficient accuracy to enable a determination to be made of facial expressions.

For this research we have focussed on AAM modelling techniques, beginning with standard AAM face models and extending our techniques to encompass specific sub-models for the eye and lip regions and eventually combining these sub-models with a global face model to develop an improved ***component-based AAM model***. When combined with a support vector decision framework, this approach demonstrates the potential of employing such component models for the robust and accurate determination of facial expression in unconstrained faces.

1.2. Summary of Principle Contributions

There are a number of main contributions arising from this research. Other secondary contributions will be described in Chapter 10.

- In Chapter 6 we present independent sub-models for both eye and lips regions of the face. In particular the eye model employs the concept of overlapping key-points which enables our model to handle open, closed, and intermediate eye-states. A patent application (currently unpublished) directed to this aspect of the model has been filed. The AAM techniques are also used to model the mouth area together with a pre-filtering technique. From our knowledge this approach was not used to independently model lips up to date.
- In Chapter 8 a component AAM model is introduced employing both a global model and local sub-models based on the eye and lips models of Chapter 6. Although component AAM models are known in the literature their use with specialised eye and lips models represents an improvement over known techniques.
- In Chapter 8 we investigate the optimal AAM parameters for modelling facial expressions and we discard the irrelevant or redundant ones. Our proposed design used the subset of model parameters which mainly encode expression information, discarding thus the information related to variability in pose and identity, leading to

more accurate expression recognition rates. Similar approach was proposed by us in [8], adapted though for face recognition use.

1.3. Thesis Outline

The thesis is structured in ten chapters starting with the introduction and concluding with the final chapter dedicated to conclusions and future directions of our research. Additionally there are four appendices included dedicated to algorithms and databases related to our work, and to several uses of our models.

Chapter 1 provides introductory information about our field of research and states the principal goals and hypothesis of this thesis. Novelties and main contributions are also described.

Chapter 2 is foundation art providing a concise introduction to a range of face modelling and analysis techniques. The goal of this chapter is to provide the reader with the necessary background knowledge to understand in detail the models and associated experiments presented in the later chapters of this thesis. Challenges that have to be addressed in order to build robust face expression recognition systems are also detailed.

Subsequent to our study, the Statistical Model of Appearance together with its search algorithm, Active Appearance Model (AAM), were determined to be the most suitable technique for our applications in view of the potential to optimise such models for use in embedded system and devices.

In ***Chapter*** 3 more specialised foundation art on the fundamentals of statistical models is introduced. The aim of Chapter 3 is to provide a background in AAM techniques to the reader. We cover the standard formulation of AAM and its state-of-the-art extensions.

Firstly, we look into the construction of AAM. Each step of the technique is thoroughly explained. In addition we discuss some of the challenges and weaknesses of AAM techniques. We also consider some of the numerous extensions to AAM which are proposed in the literature to overcome these weaknesses. We conclude this

chapter by explaining the algorithm limitations and by summarising a number of the more important research references on the AAM state-of-the-art.

The purpose of ***Chapter 4*** is to review published work that uses AAM in expression modelling, classification, and recognition. We search to understand to what extent is expression recognition influenced by gender, familiarity, symmetry, or expression intensity. This study helps us to better understand facial expression analysis and to identify its challenges. Particular attention to research which relates facial expressions to the underlying emotional states of a person is paid.

Then, the applicability of the method in the expression analysis field is surveyed. Various AAM extensions proposed in literature for still images are summarised. We searched partly to emphasise the novelty of certain techniques proposed in our work, by comparing it with the existing literature. Then, we survey expression classification and recognition based on AAM and compare results for different types of classifiers. We also describe research databases designed to provide facial images exhibiting a comprehensive range of expressions.

The main goal of our work is to be able to automatically analyse facial expressions. This task requires detailed facial feature models that accurately describe facial expressions. In ***Chapter 5***, we adapt the AAM technique in order to develop robust models for the eyes and lips regions of the face as these are the most significant facial sub-regions which determine facial expression.

We begin our approach by verifying the performance of the standard AAM formulation when modelling eyes and lips. We then adapt the AAM in order to model these challenging facial features. The eye model should depict not only the eye appearance, but also states like the degree of closure of the eye or the angle of eye gaze. It would also be important to model the state of left and right eyes independently. Finally, a lips model should be able to handle challenges, such as the complex geometric variability of the lips and the varying textures of this region as the mouth opens and a person's teeth are exposed. Then we test our initial AAM facial feature models, analysing their limitations.

In ***Chapter 6***, we propose extensions of our facial feature models, in order to solve the challenges of the standard AAM formulation and to make them robust to unseen facial variation.

In the first part of the chapter, we propose a component-based AAM adapted for the eye region. We then test the proposed model by means of some applications, such as an eye tracker and a blink detector. The second part of Chapter 6 is dedicated to an improved AAM lip model. We propose a hue filter along the AAM approach in order to obtain a better lip location. We test the lip model performance by means of some applications, such as a lip tracker and a smile detector.

In ***Chapter 7*** detailed testing and comparisons of our modelling approaches are analysed. Their performance is analysed by comparing them with standard AAMs or by testing them in the context of several consumer applications.

In ***Chapter 8*** we build on our earlier work to analyse the feasibility of integrating AAM techniques into an automatated facial expression analysis system. One of the key challenges of facial expression analysis is the extraction of relevant facial features. Our earlier work shows that AAM techniques are useful for analysis of both eye and mouth regions. Here we provide details of a comprehensive set of experiments on facial expression classification and recognition for still images and we propose a component-based AAM facial representation. We proceed to test the robustness of this framework to the deformations of typical unconstrained facial expressions.

Although the results presented at the end of Chapter 8 are very promising it is clear that additional improvements are desirable in order to achieve a truly robust and reliable facial expression analyser.

In ***Chapter 9*** we propose a series of model refinements, to further enhance our face model. The first topic we consider is the potential to use a person-specific AAM. Such a model eliminates the variability between persons from the statistical model, enabling it to better model other sources of variability. Then, we describe how our approaches for still images can be extended to track and recognise expressions in a sequence of images using person-specific AAMs. Alternatives to our basic classifiers

are proposed with a view to improve expression analysis rates, such as gender, age, and racial classification additional criteria.

Finally, the conclusions are summarised in ***Chapter 10***. Envisaged topics for further research are also suggested in this chapter.

Chapter 2 - Face Analysis Techniques

This chapter provides a concise introduction to a range of face modelling and analysis techniques. The goal of this chapter is to provide the reader with the necessary background knowledge to understand in detail the models and associated experiments presented in the later chapters of this thesis.

We begin, in section 2.1 with an overview of the main components of a recognition system for facial expressions. In brief this system takes an image as input; said image is (i) scanned by a fast face detector; (ii) a more sophisticated face model is applied to any detected face regions, enabling the extraction of detailed facial features; (iii) these features are subsequently processed to enable a recognition or verification output. An overview of the face modelling module which is critical to accurate feature extraction is given in section 2.2. Three different techniques are studied and compared: Eigenfaces (section 2.2.1), deformable models (section 2.2.2), and 3-D techniques (section 2.2.3). Based on our study, we chose the most promising face modelling technique that best suited our end-goal of a system which might prove suitable for incorporation into an embedded imaging device.

A direct application of face modelling, which developed into the core of our research and forms the central theme of this thesis, is that of facial expression recognition. Section 2.3 defines the main concepts of facial expression modelling and provides an overview of existing techniques in facial expression recognition and the challenges that remain.

In section 2.4, other face analysis techniques related to our work are described. We outline the problem of face detection in unconstrained images and focus on the Viola-Jones algorithm which is regarded as state-of-art in face detection. This particular algorithm is used to determine a starting reference for our more sophisticated models, because of its reliable performance and accuracy. Then, similarities between face recognition and facial expression recognition are analysed. We briefly compare these two applications as they present similar challenges and share many common approaches.

2.1. System Overview

A system that performs automatic face recognition or expression recognition typically is comprised of three main subsystems, as shown in Figure 2.1: (i) a face detection module, (ii) a feature extraction module, and (iii) a classification module which determines a similarity between the set of extracted features and a library of reference feature sets. Other filters or data pre-processing modules can be used between these main modules to improve the detection, feature extraction, or classification results.

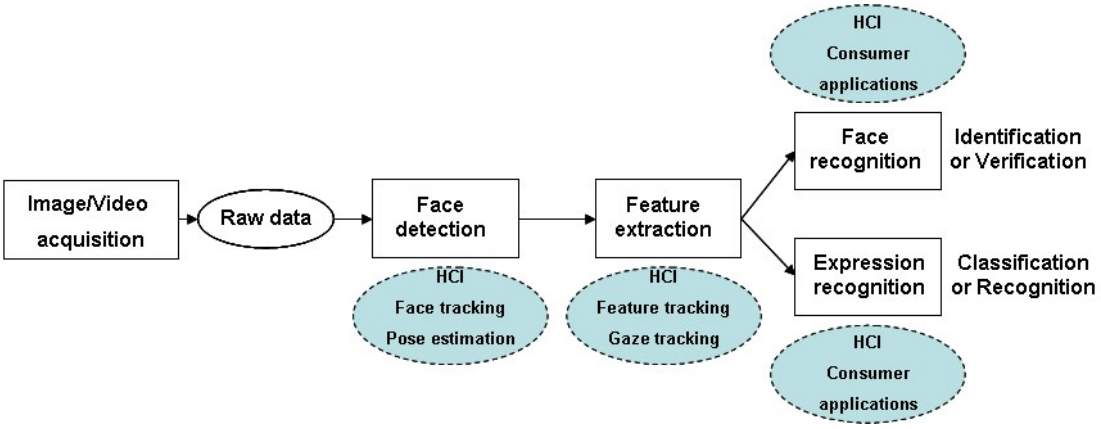


Figure 2.1. Steps to perform face recognition/facial expression recognition.

In the first module, it is decided whether the input picture or the input video sequence contains one or more faces. If so, then facial features are extracted from the detected faces, in our case by applying an advanced face model which encodes the facial features by a set of parameters. As a third step, the facial features – determined as a set of output parameters from the model - are classified in order to perform facial recognition or expression classification.

A more detailed overview of each module will be given in the remaining of this chapter.

2.2. Face Modelling

Faces are a rich source of information describing a person's identity and behaviour. Therefore, faces are complex entities, exhibiting a great deal of variation in their appearance, i.e., shape and texture.

One of the key decisions in face modelling is how to represent a face in order to include a large range of viewing conditions? Human faces have been extensively studied in vision and graphics for a wide range of tasks from detection, tracking,

expression, animation, to non-photorealistic rendering, i.e., portrait and sketch. The selection of its representation and model depends on three factors: the objectives of the task at hand, the required precision, and the resolution of the observable face images.

In the next sections we are going to define and survey a subset of face modelling approaches, concluding of the remaining challenges in this field.

2.2.1. Definition and Challenges in Face Modelling

Facial feature extraction can be viewed as the task of finding some conceptual and relevant facial information from raw data, in our case the detected face. ***Face modelling*** deals with coding these facial features by a set of parameters. The coding is useful in order to reduce the quantity of facial information to transmit, usually over a channel with low capacity. The parameters should be decoded and synthesised in the receiver side of the system with reasonable image quality.

While humans are able to detect and localise faces almost effortlessly, an automatic system faces a large number of challenges that must be dealt with, in order to achieve a good accuracy [9-17]. These challenges are due to various factors that affect the way a face is perceived. We are going to enumerate the main challenges and cite solutions proposed in literature in order to solve these challenges.

In-plane and Out-of-plane face orientation

A face can appear in a wide range of out-of-plane, e.g., frontal, 45 degrees, profile, or in-plane rotations. Therefore a face appears in many different shapes in an image. The in-plane orientation is usually solved by a shape rotation comprised in the similarity transformation (section 3.3.1), when modelling faces [18]. For the out-of-plane issue, many solutions have been proposed, such as estimating the pose through a correlation model [11] or using a full 3-D face model [10].

Figure 2.2. Faces in different poses.

Illumination conditions

The appearance of a face can change dramatically as the illumination conditions change. Due to the 3-D aspect of the face, direct illumination source can cast strong shadows and shading which affect certain facial features. It is well known that the variation due to illumination changes can be greater than the variation between individual faces. When side illumination is used, a part of the face is very bright while the other part is very dark. This can cause the occlusion of the darkest part of the face. Various methods have been proposed to overcome the illumination challenge. Examples include approximation of the human face surface with a Lambertian surface and computing a basis for a 3-D illumination subspace [9] or a geodesic illumination basis model [13].

Figure 2.3. The appearance of a face can change dramatically as the lighting conditions change.

Facial expression

Face appearance is directly affected by a person's facial expression. The appearance of a person who is laughing is significantly different from the appearance of the same person when is angry. In order to model faces independent of facial expression, researchers analyse geometric features [15], appearance features [14], or hybrids of the two types of features [17]. A more detailed discussion can be found in section 2.3.3.

Figure 2.4. Variation in face appearance caused by face expressions.

Occlusions

A face may be partially occluded by other objects or by itself, i.e., self-occlusion. A source of self-occlusion is the complex 3-D shape of the face itself. Elements partially

occluding faces can be a person standing in front of another or an object that is placed in front of the face. Also beards, moustaches, glasses, or scarves can cover facial features. Possible solutions to model occluded faces are including occluded faces in the training data [12] or using a Bayesian occluded model [16].

Figure 2.5. Examples of occluded faces.

2.2.2. Main Approaches in Face Modelling

Face modelling can be performed ***locally***, i.e. using local regions of salient features, or as a ***global*** approach, i.e., using a whole face template. We now present a short review of the most significant work described in the literature.

Since the 1990s, with the introduction of the Eigenfaces approach [19], global methods have dominated face analysis research. Compared to the abundance of global methods, e.g., [19-23], only a few techniques have been proposed to perform local appearance-based face modelling, e.g., [24, 25].

Comparisons between the two approaches have been made in the literature [26, 27]. From these comparisons, it appears that local approaches are generally superior to global approaches. Many of the sources of variability in facial appearance (listed in section 2.2.1) tend to cause great variation at global level between faces, where in a local representation scheme a change affects only the corresponding part of the representation and does not modify the facial representation as a whole.

Despite this limitation, face modelling research has been focused mainly on global approaches [19-23].

Local-feature based methods, as they are related to local regions of salient facial features, require a robust initialisation of these facial features. It is a non-trivial task to develop a precise eye/mouth detector [28, 29]. If the feature detection is not highly accurate, the error will propagate further to facial modelling and it will significantly affect the overall face modelling.

In the next sections, three global face modelling techniques will be analysed. We search for a technique that permits us to develop a detailed face model, capable of

detecting and measuring facial expression variations and that is robust to environmental changes, e.g., pose, illumination.

2.2.2.1. Eigenfaces

Eigenface [19], is commonly considered as the baseline technique for face modelling, particularly in the field of person recognition. It is based on Principal Components Analysis (PCA) techniques [30]. The Eigenface approach aims to find the principal components of a distribution of a set of face images, or the eigenvectors of the covariance matrix of a set of face images. These eigenvectors can be thought of as a set of features which together characterise the variation between face images. As the first few 2D eigenvectors retain the global shape and appearance of a face region they are known as ***eigenfaces*** (see Figure 2.6).

Figure 2.6. Eigenfaces, as presented in [19]: the average face on the left, followed by seven top eigenfaces.

Every face image is projected, after subtracting the mean face, into the principal subspace. The coefficients of the PCA expansion are averaged for each subject, resulting in a single representation of that subject. So, each face image in the training set can be reconstructed exactly as a linear combination of the eigenfaces determined from the PCA.

Face recognition, i.e., the ability to recognise people by their facial characteristics, is the source of motivation behind the creation of eigenfaces [19].

2.2.2.2. Deformable Models: Definitions and Main Approaches

Among global approaches, ***deformable models*** offer a unique and powerful approach to image analysis that combines geometry, physics, and approximation theory. They exploit bottom-up constraints derived from the image data together with top-down a priori knowledge about the location, size, and shape of the object being modelled.

A number of deformable modelling techniques which are specifically targeted at the human face have evolved. These include shape and/or appearance models, e.g., Snakes, Active Shape Model (ASM), Active Appearance Model (AAM), and 3-D

Morphable Models (3DMM). Each of these particular techniques will be described in detail and considered in the context of our research in the remainder of this section.

Deformable modelling approaches

In general, deformable models can be divided into two categories: implicit models and explicit models, or more exactly free-form models and parametric models.

Free-form models can represent any arbitrary shapes as long as some general regularisation constraint, such as continuity or smoothness, is satisfied. They are also called ***Active Contours***.

Parametric models are capable of encoding a specific characteristic shape and its variation. The shape is characterised by a parametric formula and by modes of deformation. For a detailed review, please refer to the survey paper by [31].

A popular and well-known class of deformable model is that of ***Snakes*** or ***deformable contour*** models [23]. Snakes are planar deformable contours that are useful in several image analysis tasks. They are often used to approximate the locations and shapes of object boundaries in images based on the reasonable assumption that boundaries are piecewise continuous or smooth. Snakes can deform to match any smooth contour. They are best suited as an aid in manual segmentation, since they have no a priori knowledge of the domain.

A smart Snake algorithm, as it has been called, is the ***Active Shape Model (ASM)*** [20]. ASM is a statistical model describing the shape of an object which iteratively deforms to fit to an example of the object in a new image. The shapes are constrained by the Point Distribution Model (PDM) [22] to vary only in ways seen in a training set of labelled examples. The model is able to capture the natural variability within a class of shapes. It includes this large range of variability, but it is still specific to the class it represents.

The key difference between Snakes and ASM is that the latter can only deform in ways consistent with the data in the training set. The ASM fitting algorithm aims to match the model to a new image. It iteratively looks in the image around each point for a better position for that point and it updates the model parameters to best match these new found positions.

A direct extension of the ASM approach has lead to the ***Active Appearance Models (AAM)***. In ASM, the feature model is 1-D profile texture feature around every feature point. In AAM, the global appearance model is introduced to conduct the

optimisation of shape parameters. AAM is a generalisation of ASM, but uses all the information in the image region covered by the target object, rather than just that near modelled edges. Generally speaking, ASM outperforms AAM in shape localisation accuracy and it is more robust to illumination, but has local minima problems. AAM is sensitive to illumination and a noisy background, but can get optimal global texture. In essence AAM can be seen as a successful combination between the ideas of Eigenfaces [19], PDM [22], and ASM [20], with some ingenious refinements.

2.2.2.3. 3-D Models

3-D modelling of data in pattern recognition tasks has been largely considered as a means to overcome variations in pose and illumination [32]. The 3-D facial data can provide more geometric information for face modelling than that of 2-D images. With the development of 3-D acquisition system, 3-D capture has become significantly faster and cheaper, and techniques based on 3-D information are attracting much attention [33].

One of the most common 3-D approaches is called ***Morphable Modelling (3DMM)*** [2]. Blanz and Vetter derive the 3DMM by transforming the shape and texture of 3-D face examples into a vector space representation. New faces and expressions can be modelled by linearly combining face prototypes or by varying attributes of a single face (see Figure 2.7). In their method, 3-D faces can either be generated automatically from one or more photographs, as in [34], or modelled directly by a set of 3-D range scans.

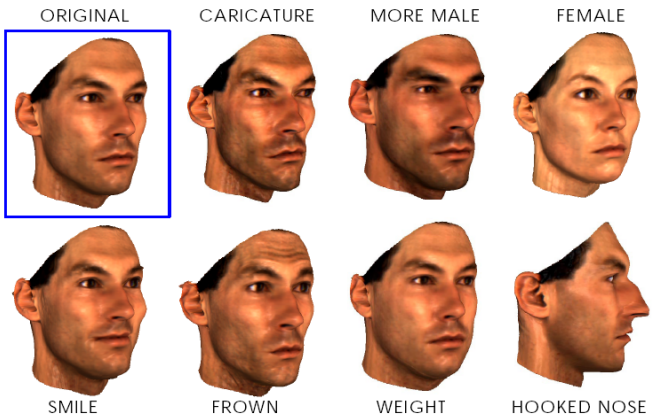


Figure 2.7. 3-D variation of facial attributes of a single face [10]. The appearance of an original face can be changed by adding or subtracting shape and texture vectors specific to the attribute.

2.2.3. Comparison of Facial Models

It is almost two decades since Eigenfaces [19] first came to the attention of the computer vision research community. Since its introduction, many other 2-D or 3-D techniques or extensions of the Eigenfaces [35] have been proposed. Two other important techniques are presented in section 2.2.2.2, and, respectively, 2.2.2.3.

Since their introduction, 2-D deformable models became popular among researchers. The deformable models adapt themselves to fit the given data. They are more powerful than rigid models, due to their capability to deal with shape deformations and variations.

A comparison between the 2-D and 3-D face models is performed in [36] along three axes: representational power, construction, and real-time fitting. The authors conclude that, overall, 3-D models are preferable to 2-D models. From their experiments, they demonstrated that the 3-D parameterisation is more compact, i.e., up to six times fewer parameters, and more natural, i.e., head pose and non-rigid shape deformation are separated. 3-D model fitting is more robust and requires fewer iterations to converge; also 3-D occlusion reasoning is more powerful.

In the same study [36], it is nevertheless concluded that, in practice, the differences between 2-D and 3-D models are not as great as is frequently claimed. Ignoring the size of the model and assuming a scaled orthographic camera, the 2-D models have the same representational power as 3-D models. Separating the head pose from the non-rigid head shape deformation in 2-D is possible, and approximate occlusion reasoning is achievable.

The one limitation of 2-D models that is hard to avoid is that their fitting is inherently less robust than of the 3-D models [36].

However, a significant negative aspect of a 3-D representation is that it requires a 3-D acquisition system or methodology to obtain 3-D data from 2-D data [34]. Also few 3-D face databases [37] are publicly available and the number of subjects in existing databases is limited, and the quality is somewhat inconsistent.

Now, the choice of the face modelling technique is taken in function of its suitability for the type of the applications desired by the user. As explained in the introductory chapter our research is funded by FotoNation (Ireland) and its parent corporation, Tessera Inc. As a consequence our research was targeted to investigate

approaches which have potential value for consumer electronic devices, e.g., we wish to embed our face models in smart digital cameras for various applications, such as face recognition or facial expression recognition.

AAM is recommended by its wide applicability, compact representation of variations of the model, speed, accuracy, and robustness. Another main advantage of AAM, when compared to other 2-D techniques, is that it removes noise [36]. In the same paper [36] it is stated that the noise removal in the AAM process and its ability to use far more data in the training process, dramatically improve a system performance. Also, AAM can be relatively easily implemented in consumer applications [38], hence our choice to use AAM as a basis for the research activities described in the remainder of this thesis.

In the next section, we present an overview of facial expression modelling techniques and, in particular, we consider the suitability of AAM to model facial expressions.

2.3. Facial Expression Modelling

In the previous section we explained some of the complexities of human facial appearance and surveyed techniques of face modelling. Complexity in facial appearance can be caused, among other factors, by facial expression variations. It is well-known that even the simplest of facial expressions, e.g. a wink, creates many changes in the facial pattern. AAM was chosen to model facial appearance. Is this technique also suitable to model facial expression variations?

In the current section we define terms related to facial expressions (section 2.3.1), discuss challenges in modelling facial expression variations (section 2.3.2), and survey facial expression modelling techniques (section 2.3.3). Subsequent to our research, we chose AAM as the most appropriate technique for our face expression analysis system and for our applications (section 2.3.4).

2.3.1. Definitions of Facial Expression Related Terms

It is important to emphasise from the beginning that there is a distinction from a computer vision point of view between facial expression recognition and human emotion recognition. As Fasel and Luettin explain [39], the former deals with the classification of facial motion and facial feature deformation into abstract classes

based on visual information. The latter is a result of many different factors and it can be revealed on more channels, e.g., voice, pose, gestures, gaze direction, and facial expression. To match a facial expression with an emotion implies knowledge of the categories of human emotions into which expressions can be assigned.

Researchers focused their attention on a coding system for facial expressions. ***Facial Action Coding System (FACS)***, originally developed by Ekman and Friesen in 1976 [40], is the most widely used coding system in the behavioural sciences. The system was originally developed by analysing video footage of a range of individuals and associating facial appearance changes with contractions of the underlying muscles.

The outcome was an encoding of 44 distinct ***action units (AUs)***, i.e., anatomically related to contraction of specific facial muscles, each of which is intrinsically related to a small set of localised muscular activations. Using FACS, one can manually code nearly any anatomically possible facial expression, decomposing it into the specific AUs and their temporal segments that produced the expression. All resulting expressions can be described using the 44 AUs described by Ekman or a combination of the 44 AUs. In 2002, a new version of FACS is published, with large contributions by Joseph Hager [41].

Ekman and Friesen [42] have also postulated six primary emotions which they consider to be universal across human ethnicities and cultures. These ***six universal emotions***, commonly referred as ***basic emotions*** are: happiness, anger, surprise, disgust, fear, and sadness (Figure 2.8). The leading study of Ekman and Friesen [42] formed the origin of facial expression analysis, when the authors proclaimed that the six basic prototypical facial expressions are recognised universally. Most of the researchers argue that these expressions categories are not sufficient to describe all facial expressions in detail. However, most of the existing facial expression analysers still use Ekman and Friesen's theory.

Figure 2.8. On each row are represented the six universal facial expressions (in order, anger, happiness, neutral, surprise, fear, sadness, and disgust) and the neutral state, as expressed by different subjects.

There is a connection between FACS and the basic emotions. Each basic emotion was described by Ekman and Friesen [43] using specific cues describing facial activity. This description is summarised in Table 2.1.

In our work, we will study expression recognition using these basic emotions categories. The choice of decoding the six universal emotions is based on the straightforwardness of the universal emotions compared to FACS. Also, the analysis of the universal emotions gives us the possibility to evaluate our work versus the work of other researchers who use the same framework.

In our work, we consider each of a particular class of facial expressions as the source or direct representation of the equivalent human emotional state. Thus the terms “facial expression” and “facial emotion” are used interchangeably throughout the remainder of this thesis.

TABLE 2.1. CUES FOR FACIAL EMOTIONS DETAILED BY EKMAN AND FRIESEN [43].

Emotion	Observed facial cues
Surprise	Brows raised (curved and high)
	Skin below brow stretched
	Horizontal wrinkles across forehead
	Eyelids opened and more of the white of the eye is visible
	Jaw drops open without tension or stretching of the mouth
Fear	Brows raised and drawn together
	Forehead wrinkles drawn to the centre
	Upper eyelid is raised and lower eyelid is drawn up
	Mouth is open
	Lips are slightly tensed or stretched and drawn back
Disgust	Upper lip is raised
	Lower lips is raised and pushed up to upper lip or is lowered and slightly protruding
	Nose is wrinkled
	Cheeks are raised
	Lines below the lower lid, lid is pushed up, but not tensed
	Brows are lowered, lowering the upper lid
Anger	Brows lowered and drawn together
	Vertical lines appear between brows
	Lower lid is tensed and may or may not be raised
	Upper lid is tensed and may or may not be lowered due to brows' action
	Eyes have a hard stare and may have a bulging appearance
	Lips are either pressed firmly together with corners straight or down, or are open, tense in a squarish shape
	Nostrils may be dilated (could occur in sadness too)
	Unambiguous only if registered in all three facial areas
Happiness	Corners of lips are drawn back and up
	Mouth may or may not be parted with teeth exposed or not
	A wrinkle runs down from the nose to the outer edge beyond lip corners
	Cheeks are raised
	Lower eyelid shows wrinkles below it and may be raised but not tense
	Crow's-feet wrinkles go outward from the outer corners of the eyes
Sadness	Inner corners of eyebrows are drawn up
	Skin below the eyebrow is triangulated, with inner corner up
	Upper lid inner corner is raised
	Corners of the lips are drawn or lip is trembling

2.3.2. Challenges in Modelling Facial Expression Variations

Although humans recognise facial expressions without effort or delay, reliable facial expression recognition by machine is still a challenge. Face variability (as described in section 2.2.1) is the main challenge and the best way to improve expression modelling is to control this variability; however if we deal with unconstrained images such control is not available.

In the next section we survey existing techniques for modelling facial expressions and consider how such techniques may be further refined to enhance their robustness to variability in the input face regions.

2.3.3. Main Approaches of Facial Expression Modelling

The facial expression modelling approaches differ from the point of view of the employed feature extraction method and of types of facial features used for modelling.

In this section, we describe different types of facial features and methods to extract facial features. In the next stages of expression recognition, the extracted facial features are classified in order to perform facial expression classification/ recognition. More details about expression classifiers will be given in Chapter 4.

2.3.3.1. Facial Features

Several facial feature categories can be extracted, e.g., geometric features [15, 44-46], texture features [14, 47-50], or hybrid combinations of the two [17, 47, 51, 52]. We next explain what is meant by these.

Geometric Features

Geometric features describe the shape of facial components, i.e., mouth, eyes, eyebrows, nose, and their locations, i.e. corners of the eyes, corners of the mouth, etc. They are represented by facial components or facial feature points, forming a feature vector that represents face geometry. Geometric feature extraction can be computationally more expensive, but it is more robust to variation in face position, scale, size, and head orientation.

Pantic and Rothkrantz [15] successfully employ this type of features. In their research, a point-based model is described by a set of facial characteristic points (see Figure 2.9). An important contribution is the fact that the model is composed of two 2-D facial views, namely the frontal and the side view. They are considered separately and they do not contain redundant information about facial features. When coupled together, however, the two facial views reveal redundant information about facial expression.

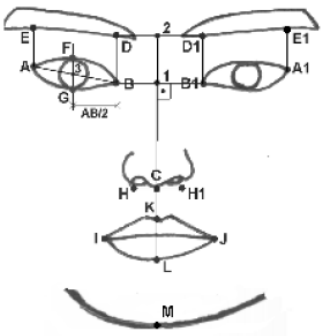


Figure 2.9. Facial points of the frontal-view [15].

Later, Gokturk et al. [44] used a nineteen point face mesh to introduce a 3-D model-based tracker. The tracker extracts simultaneously and robustly the pose and the shape of the face at every frame of a video sequence. Pose and shape characteristics are naturally factored into two separated signatures, leading to good facial expression classification rates, even under extreme head pose configurations.

Chang et al. [45], using a shape model defined by fifty-eight facial landmarks, learn a low-dimensional manifold with a subsequent probabilistic modelling used for expression tracking and recognition.

Moreover, Kotsia et al. [46] exploit geometrical features using a Candide model. The user has to manually place some of Candide grid nodes to face landmarks depicted at the first frame of the image sequence. The grid-tracking-and-deformation system, based on deformable models, tracks the grid in consecutive video frames (see Figure 2.10). Geometrical displacements of certain Candide nodes are extracted.

Figure 2.10. Seven Candide nodes are fitted on a facial image [46].

Texture Features

Texture features are also often called appearance features. In our thesis, we will use the term of texture, as appearance is used to define both shape and texture. Texture features describe the appearance, i.e., skin texture, and the changes of the face, i.e., wrinkles, furrows, or bulges. The texture features can be extracted on either the whole-face or specific regions in a face image.

One approach to texture-feature extraction is performed using Gabor wavelets [14, 47]. In [14], specific locations of the face are coded using a set of Gabor filters topographically ordered and aligned approximately with the face.

Zhang et al. [47] also adopt this method. They compare the expression modelling performance of geometric-features, i.e., the positions of thirty-four fiducial points on a face, and texture-features, i.e., six hundred and twelve Gabor wavelet coefficients extracted from the face image at these thirty-four fiducial points. Better results are obtained from the Gabor coefficients.

Silapachote et al. use the AdaBoost approach to extract texture features relevant for facial expressions [48]. They select the relevant global and local texture features by choosing the ones with the most discriminating information. Their selectivity reduces the dimensionality of the feature space, resulting in significant speed gains during online classification.

Bartlett and her colleagues [50] use eigenfaces as texture features for expression classification. Three approaches are compared: holistic spatial analysis, explicit measurement of features such as wrinkles, and estimation of motion flow fields. Best results are obtained when the three methods are combined in a hybrid system that successfully classifies expressions.

Hybrid Features

The hybrid features are described by both geometric and texture features. Several researchers use the hybrid features to improve expression decoding accuracy.

Tian et al. [17] use hybrid features in order to analyse facial expressions, based on both permanent facial features, i.e., brows, eyes, mouth, and transient facial features, i.e., deepening of facial furrows. Multi-state face and facial component models are proposed for tracking and modelling facial features.

Wen et al. [51] present an adaptive technique for analysing subtle facial appearance changes, including also facial feature point motion. They use a new ratio-image based appearance feature, independent of a person's face albedo.

There are several comparisons in literature between the various types of features. Research shows that appearance-based features can achieve better results than the geometric features [47, 50]. Moreover, hybrid features outperform the appearance and geometric features [53].

2.3.3.2. Feature Extraction Approaches

A number of approaches have been proposed for feature extraction. Most attempts focus on optical flow analysis [54], model-based approaches [4, 55-60], or image-based approaches [61, 62].

The ***optical flow*** approach [54] involves assessing the magnitude and the direction of facial motion. It can be used either as a global algorithm or applied on local regions. Optical flow generates a matrix of vectors originating from a pixel position in one image and terminating at a pixel position within another image. These vectors represent the mapping between the pixel intensities in one image and the pixel intensities in another. These displacement vectors may be used to describe correspondents between similar images.

Model-based approaches extract information relying on extensive knowledge about the object of interest. These approaches include PCA [60], Eigenfaces and Fisher Faces [55], ASM [58, 59], or AAM [56, 57].

Alternatively, ***image-based*** approaches, e.g., Gabor wavelet [62], extract features from images without relying on general knowledge about the object to be modelled.

2.3.4. Comparison of Approaches describing Facial Expression Variations

Comprehensive surveys reviewing facial expression analysis are provided in the literature [39, 63-65]. In [63], Pantic and Rothkrantz survey the work done in achieving accurate real-time systems that detect faces, extract facial expression information, and classify expressions. Automatic facial expression analysis is also reviewed by Fasel and Luettin [39], with an emphasis on face normalisation, facial expression dynamics and their intensities, and the robustness towards environmental changes. Recent advances in machine analysis of expressions and recent work of two leading research groups in this field, namely that of Pantic and her colleagues, and that of Bartlett and her colleagues, are summarised in [65]. A complete guide to facial expression analysis, including more recent advances, is provided by Tian et al. [64]. Various types of techniques for facial expression analysis are enumerated in these surveys. A summary of these techniques was given in section 2.3.3.

In [66], Tian et al. describe the properties of an ideal analysis system, as summarised in Figure 2.11. The authors conclude that none of the proposed methods in literature are able to entirely satisfy these requirements.

Robustness	
Rb1	Deal with subjects of different age, gender, ethnicity
Rb2	Handle lighting changes
Rb3	Handle large head motion
Rb4	Handle occlusion
Rb5	Handle different image resolution
Rb6	Recognize all possible expressions
Rb7	Recognize expressions with different intensity
Rb8	Recognize asymmetrical expressions
Rb9	Recognize spontaneous expressions
Automatic process	
Am1	Automatic face acquisition
Am2	Automatic facial feature extraction
Am3	Automatic expression recognition
Real-time process	
Rt1	Real-time face acquisition
Rt2	Real-time facial feature extraction
Rt3	Real-time expression recognition
Autonomic Process	
An1	Output recognition with confidence
An2	Adaptive to different level outputs based on input images

Figure 2.11. Properties of an ideal expression analysis system, as defined by [66].

When comparing feature extraction approaches, the authors of [66] conclude that the optical flow approach is relatively insensitive to subtle motions that are rich in

facial expression information. However it is sensitive to illumination variations, images of low quality, or non-rigid motion. Conversely, image-based techniques are fast and simple. They work well under environmental constraints, but lack in robustness towards pose, illumination, or non-rigid changes. They also require reliable facial features.

Facial expressions are defined mainly by contraction of facial muscles that produce changes in facial appearance and shape [67]. In [68] it is stated that model-based approaches have an inherent benefit over purely image-based representations, in the sense that they can account for many types of "real-world" variations of the face appearance.

As a result, this technique is known to have a potential value in facial expression applications. AAM yields good results on difficult and noisy data and it benefits from real-time model optimisation algorithms [69]. AAM can be suited to embedded system environments, such as mobile camera phones [38].

2.4. Other Face Analysis Techniques

2.4.1. Face Detection and the State-of-the-art in Face Detection

From our experience, we learned that achieving accurate initial face detection is essential to any subsequent facial modelling or analysis. In the particular example that we give in Figure 2.12, AAM techniques [56, 57] fail due to the poor initialisation. We remark that most face models will suffer similar poor convergence if an accurate initialisation of the model, both in terms of size and position, is not available.

In our system, we employ a face detector module as initialisation for the AAM search. ***Face detection*** can be defined as the ability to detect and localise faces within an image or a scene. In the past several years, many face detection techniques have been proposed in literature. A comprehensive survey of the face detection methods is presented in [70]. State-of-the-art face detection methods provide real-time solutions that report high detection rates.

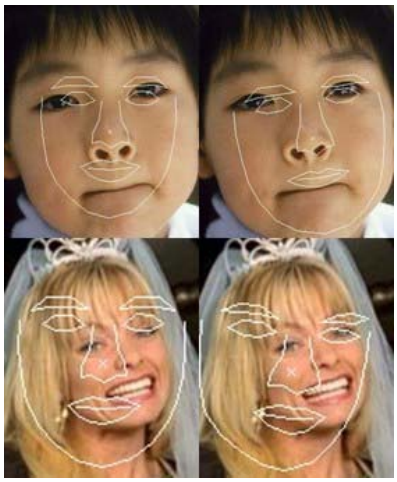


Figure 2.12. Bad AAM fitting, caused by poor initialisation (first column displays the shape after the initialisation procedure, while the second column displays the incorrect found shape after the AAM search, consequence to poor initialisation).

The main idea of the technique is to combine weak classifiers based on simple binary features, which can be computed extremely fast. Simple rectangular Haar-like features are extracted; face and non-face classification is done using a cascade of successively more complex classifiers that discards non-face regions and only sends face-like candidates to the next layer's classifier. Each layer's classifier is trained by the AdaBoost learning algorithm. AdaBoost is a boosting learning algorithm that can fuse many weak classifiers into a single more powerful classifier.

In our system, we employ the Viola-Jones face detector, as initialisation for the AAM search. An example of face detection using the Viola and Jones algorithm is shown in Figure 2.13. This algorithm has been implemented in OpenCV [71], which is the free computer vision library developed by Intel. It is this implementation that has been used in practice throughout our current work.

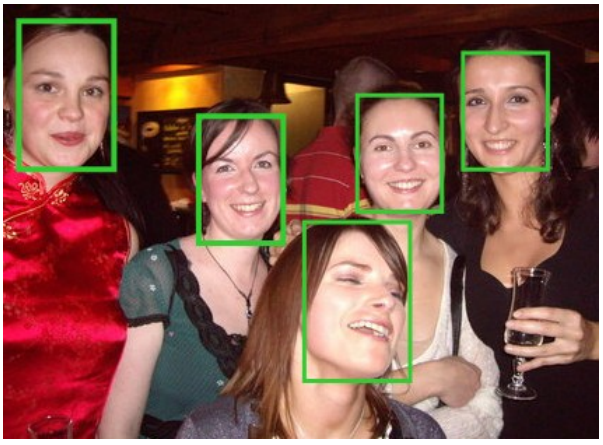


Figure 2.13. Example of detected faces using the Viola and Jones algorithm.

2.4.2. Face Recognition (Relationship and Similarities to Facial Expression Modelling)

Perhaps the most common use of face modelling is to perform face recognition. AAM has extended its applicability also in face recognition [8, 72]. In our research group, AAM applied to face recognition was investigated in parallel with our work, by a fellow PhD researcher, Mr. Mircea Ionita [8]. His work formed the basis for some of the experiments described in this thesis (such as the texture normalisation or the illumination robust algorithm summarised in Appendix D).

Technically speaking we should differentiate between several specific forms of classifying or recognising people from their faces:

Face recognition deals with the ***identification*** and the ***verification*** of one or more persons in a still frame or in a sequence of frames, using a stored database of reference faces.

Face verification, i.e., "Am I who I say I am?", is a one-to-one match that compares a query of face images against a template face image whose identity is being claimed.

Face identification, i.e., "Who am I?", is a one-to-many matching process that compares a query face image against all the template images in a face database to determine the identity of the query face.

In recent years face recognition has received substantial attention from researchers in both academic and industrial environments, in various domains such as biometrics, computer vision, pattern recognition, or machine learning. Despite of a plethora of articles and solutions proposed, face recognition continues to be a challenging task. Many face recognition systems are proposed world wide and despite their success in controlled conditions, e.g., frontal non-occluded faces, no variations in illumination, there is not yet a common adopted approach for analysing and classifying face regions inside an image.

There are numerous similarities between face recognition and facial expression recognition. Existing work in both applications has demonstrated good recognition performance in controlled conditions. The challenge for researchers is

when such ideal conditions are not satisfied. Solutions to overcome this challenge, such as the techniques mentioned in section 2.2.1, are common to both applications.

The existing face recognition and face expression recognition systems usually require very large training datasets and significant computing power in order to be able to provide faster recognition results. Also, both face recognition and face expression recognition necessitate a strong initialisation of the face and a detailed face model. As described in section 2.4.1, the applications require robust face detection and accurate face feature extraction.

Chapter 3 - Statistical Model of Appearance

In the previous chapter we reviewed face modelling and face expression modelling techniques. The Statistical Model of Appearance together with its search algorithm, Active Appearance Model (AAM), were determined to be the most suitable technique for our application in view of the potential to optimise such models for use in embedded system and devices. This thesis is structured around this particular form of face modelling technique and we will later investigate a number of enhancements of state of the art including sub-models for the eye and mouth and a component modelling technique to combine a global face model with such sub-models. First, though, we must explain the fundamentals of such statistical models.

The aim of this chapter is to provide a proper insight into relevant aspects of AAM, such as its standard formulation and its state-of-the-art extensions.

An overview of AAM is given in section 3.1. Sections 3.2 and 3.3 are dedicated to the construction of the statistical model of appearance and of its corresponding search algorithm. Each step of the technique is thoroughly explained.

Despite its success, AAM still has to overcome common face modelling challenges, as described in section 3.4. Hence, numerous extensions of AAM are proposed in literature, both regarding the model realisation or the search algorithm. Section 3.5 summarises and compares important references of the AAM state-of-the-art.

3.1. Background to Statistical Models of Appearance

A Statistical Model of Appearance together with its corresponding search algorithm, AAM, is a powerful tool for modelling images of deformable objects. The technique was originally proposed by Cootes et al. [21] in 1998 as a deformable model, capable of interpreting and synthesising new images of the object of interest. Such models are particularly effective in modelling deformable objects with limited variability, often biological shapes, e.g., human faces, human organs. AAM proved its capability in various recognition tasks, such as medical imaging [73], face interpretation and tracking [74], and face recognition [75].

The technique includes two concepts: firstly it builds its statistical models of shape and texture. This is also known as Statistical Models of Appearance. Secondly

it includes an algorithm that permits to find, analyse, and synthesise objects belonging to the statistical modelled class in unseen images. This is also known as AAM. It is commonly thought that researchers use the term AAM to express both concepts: the statistical model of appearance and the search algorithm. This convention is often used in this thesis.

AAM uses Principal Component Analysis (PCA) [30] based linear subspaces to model the shapes and textures of the images of a target object class. Such a representation allows AAM to represent a certain image with a reduced number of parameters. Given a previously unseen image that belongs to the same object class, AAM finds the optimal parameters to represent the target image by using an iterative scheme that is fast and robust.

Each step of this technique is described in further detail in the following sections.

3.2. Building a Statistical Model of Appearance

The first step in building a Statistical Model of Appearance is data acquisition. This is followed by a suitable normalisation, and then the data are analysed and modelled based on statistical analysis.

As a first step in this process, a shape model is developed - this is described in section 3.2.1; the second step is to perform texture modelling - as described in section 3.2.2.

Now these models can be used independently, i.e., concatenated into an appearance model where the model parameters independently affect shape or texture, but not both; alternatively they may be merged through an additional PCA process to provide a combined appearance model with a reduced parameter set, but where each of the model parameters affects both shape and texture. In other words we have simplified the model by using the correlation between shape and texture to reduce the number of input parameters. These two approaches, referred to as independent and combined AAM, will be discussed and compared in detail in section 3.2.3.

3.2.1. Shape Modelling

Shape is defined as that quality of a configuration of points which is invariant under some transformations. In 2-D or 3-D we usually consider the similarity

transformation, i.e., translation, rotation, and scaling. So, the shape of an object should not change when it is moved, rotated, or scaled.

Face annotation

The shape of an object is represented by a fixed set of points, i.e., ***landmarks.*** An important decision is the choice of the landmarks: their number and their position. The landmarks should describe the important details of an object, in our case be representative of the facial features. Points could be placed at clear corners of object boundaries, edges, and T junctions between boundaries, or at easily located biological landmarks. However, in the case of human faces there are not enough of such points to give anything more than a sparse description of the shape and principle features. To overcome this problem we need to place additional equi-distant points between the representative landmarks. Examples of different methods of annotating faces are given in Figure 3.1.

Figure 3.1. Different types of detailed landmarking describing faces.

At present the landmarking of a training set is semi-automated, using templates and edge detection techniques but also requiring some fine-tuning by hand. Automatic landmarking systems are under development [76] but the wide variability of faces, particularly in unconstrained environments, makes this a significant research challenge in its own right.

One common well-known semi-automated method is to annotate a small number of examples and build a model using them. The model is fitted on the remaining images in the training set, obtaining new annotations for these pictures. Then, if necessary, the points can be manually corrected. Motivated by the simplicity and accuracy of this method, we employed it in the construction of our models.

After choosing suitable landmarks to describe the desired shape, a shape vector is given by the concatenated coordinates of all landmark points. It is formally written as $s=(x_1,x_2,\ldots,x_L,y_1,y_2,\ldots,y_L)^T$, where L is the number of landmark points. Given N training examples, we generate N such vectors s. Before we can perform statistical analysis on these vectors, it is important that the represented shapes are aligned to a reference shape. Subsequently, an alignment algorithm is applied.

Shape alignment

There is considerable literature on methods of aligning shapes into a common co-ordinate frame, but the most popular approach can be considered the ***Procrustes Analysis*** [77]. It is an iterative technique used to remove translation, 2-D rotation, and scale differences between the training shapes. It is based on minimising the distances from shapes to the mean shape. The algorithm can be summarised as following:

- Each shape is translated so that its centre of gravity is at a common origin.
- One shape is chosen as an initial estimate of the mean shape and it is scaled so that $|s| = 1$.
- The first estimate is saved as s_0 and used to define the default reference frame.
- All shapes are aligned with s_0.
- Iteratively, the mean shape is re-estimated from the aligned shapes.
- Constraints are applied on the current estimate of the mean by aligning it with s_0 and scaling so that $|s| = 1$.
- If convergence is not obtained, all the shapes are re-aligned with the current s_0 and the steps are continued.
- Convergence is thus declared when the mean shape does not change significantly within an iteration. It was observed [18] that two iterations of the algorithm are sufficient in most cases.

Modelling shape variation

After obtaining an aligned set of shapes, intra-class shape variation should be described consistently and efficiently. To simplify the problem, we first wish to reduce the dimensionality of the data to something more manageable and efficient.

An effective and well-known approach is to apply PCA [30] to the data. PCA computes the main axes of the data cloud, allowing one to approximate any of the

original points using a model with fewer than the original number of parameters. The PCA algorithm is detailed in Appendix A.

The shape model is obtained by applying PCA on the set of aligned shapes:

$$s = \bar{s} + \varphi_s b_s \tag{3.1}$$

where $\bar{s} = \frac{1}{N_S}\sum_{i=1}^{N_S} s_i$ is the mean shape vector, and N_S is the number of shape observations; φ_s is the matrix having the eigenvectors as its columns; b_s defines the set of parameters of the shape model.

By varying the elements of b_s we can vary the shape *s,* described in Equation (3.1). The variance of the i^{th} parameter, b_i, across the training set is given by λ_i. By applying limits of $+/-3\lambda_i$ to the parameter b_i, we ensure that the generated shape is similar to those in the original training set.

The allowed variance can be chosen so that the model represents some proportion of the total variance of the data; the residual terms can be considered noise. Typically a value of 0.95 to 0.98 is chosen, i.e. the original data can be reconstructed within, say, a 98% accuracy. In a description of how a statistical model of shape models the geometry of a face [78], an elegant example of a linear combination of shape parameters and shapes depicting an unseen shape is presented in Figure 3.2:

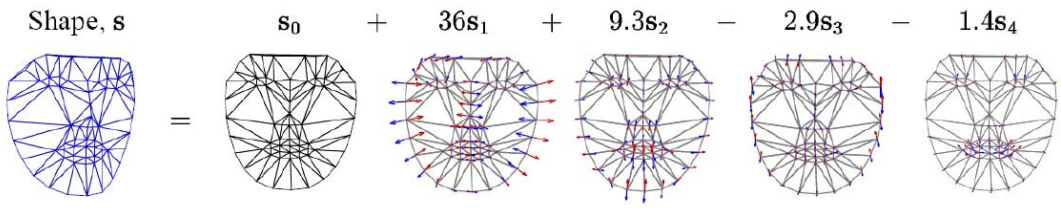


Figure 3.2. An example of the combination of the 2-D AAMs shape models, as shown in [78], that creates an unseen shape, s.

3.2.2. Texture Modelling

To complete the appearance of the model, we have to take into account also its texture. In this technique, the ***texture*** is defined as the variation in pixel values across the object of interest [79]. It is statistically modelled following a similar algorithm as for the shape.

However, in the case of texture a consistent method of collecting the texture information from the 2-D regions delineated by the landmarks is required. Therefore, a warping transform is used to project the pixel intensities across each shape from the annotated training images into a reference shape in order to generate matching texture vectors. This reference shape is normally chosen as the mean shape determined from the training set as described in section 3.2.1.

Warping stage

Image warping is a geometric image transformation used to match two images by finding an optimal spatial and pixel value transformation which best aligns them with respect to certain criteria. Image warping from image *I* to image *I'* can be defined as:

$$I'(x,y) = T(I(W(x,y))), \tag{3.2}$$

where *W* is a 2-D spatial coordinate transformation and *T* is a transformation in the texture appearance.

The choice of the warp technique is based on a compromise between a technique that generates a smooth distortion and one which achieves a good match [78]. For a survey on warping techniques refer to Glasbey and Mardia [78].

Since AAMs are landmark-based it is common practice to use the class of image warping techniques that considers the mapping of one arbitrary point set into another. A face patch is warped into the mean shape using a ***triangulation algorithm***, so that its control points match the mean shape. The triangulation algorithm is used to partition the convex hull of the control points describing the shape.

In the original AAM formulation dense image correspondences were established by image sampling via a shape-normalised, or shape-free, triangular mesh [80]. Fixed sampling points on this mesh were propagated using Barycentric coordinates to a mesh similar in structure, but deformed in shape. However, warp degeneracy can arise if triangle normals are inverted, e.g. if two vertices of a triangle stay fixed and the third is mirrored. Fortunately, for high-quality meshes and moderate shape deformations this will rarely happen. To this end, triangular meshes are typically established using a ***Delaunay triangulation***.

Delaunay triangulation technique connects an irregular point set by a mesh of angles satisfying the ***Delaunay property***. This means that no triangle has any points

inside its circumcircle, which is the unique circle that contains all three corners, i.e., vertices, of the triangle. Examples of faces warped into a mean shape and the triangulation of a face are shown in Figure 3.3.

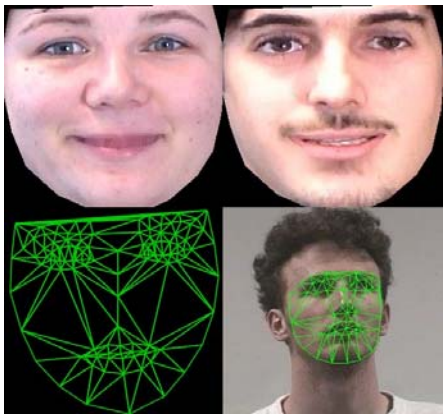


Figure 3.3. Examples of free-shape face image and the corresponding triangulation of a face.

The points inside triangles are then mapped via an affine transformation which uniquely assigns the corners of a triangle to their new positions. AAM typically only uses 50-100 mesh vertices, as they are usually constructed from a set of training images with the vertices hand-labelled on them. In practice, AAM meshes are best defined on the few facial landmarks that are easy to locate.

The Delaunay triangulation is the most popular technique used by researchers because of its simplicity. It was proven to work well for frontal faces and where there are small occlusions. We also employ this technique in the development of our models.

Now there are some drawbacks of this technique that led to the development of alternative techniques [81]. These drawbacks appear in particular when we deal with a large number of landmarks describing the shape. In these cases, when large pose variations of faces or occlusions occur, then the corners of some triangles tend to get reversed causing the inversion of the triangle normals as described above.

A further drawback is the fact that this transformation does not guarantee that straight lines across triangles are being preserved. In this case “false” objects can be created. The errors are further propagated into the search algorithm, resulting in an incorrect fitting. More advanced and accurate techniques, such as Thin Plate Splines (TPS) [81] can be employed to solve these drawbacks [82].

Nevertheless we did not encounter any significant incidence of these problems in our work and we concluded that such problems are very infrequent for the number of landmarks chosen for our models.

It is also worth noting that the rationale to use more landmark points is to achieve a more accurate face model. Our research into component models demonstrates that such an increase in modelling accuracy can be better achieved using fewer landmark points, but integrating these within a framework of several complimentary AAM models.

Texture Normalisation

To minimise the effect of global lighting variation, faces are normalised. This is a very important stage in the model creation, as illumination variations have a severe negative effect on face texture. The training samples are normalised by applying a scaling α and offset β, as follows:

$$t = \frac{t_{im} - \beta 1}{\alpha} \tag{3.3}$$

where *1* is the unity vector. α is a scaling factor, while β is the offset.

The values of α and β are chosen to best match the vector to the normalised mean. In practice, the mean normalised texture vector is offset and scaled to have zero-mean and unit-variance. Obtaining the mean of the normalised data is a recursive process, as the normalisation is defined in terms of the mean. A stable solution can be found by using one of the examples as the first estimate of the mean, aligning the others to it, re-estimating the mean and iterating.

Modelling texture variation

A texture vector $t=(t_1,t_2,...,t_p)^T$ is built for each training image by sampling the values across the warped samples, i.e. shape normalised patches, with p being the number of texture samples.

The texture model is also derived by means of PCA on the texture vectors:

$$t = \bar{t} + \varphi_t b_t \tag{3.4}$$

where $\bar{t} = \frac{1}{N}\sum_{i=1}^{N_t} t_i$ is the mean texture vector, with N_t as the number of texture observations; φ_t is the matrix of eigenvectors, and b_t the texture parameters.

As in the shape model case, it is common to allow between 95% and 98% variance of the model the lower value allowing greater flexibility in fitting the model to unseen faces.

3.2.3. The Combined Statistical Model of Appearance

The ***appearance*** of objects is described by shape and texture parameters. There are two approaches to unifying these parameters: they can be used either ***independently***, concatenated in an appearance vector, or ***combined*** by a third PCA process. In here, PCA has the role of remove correlation between shape and texture model parameters and to make the model representation more compact.

In the first approach, the sets of shape and texture parameters are concatenated as follows:

$$c = \begin{pmatrix} W_s b_{s_{opt}} \\ b_{t_{opt}} \end{pmatrix} \tag{3.5}$$

where *Ws* is a suitable weighting between pixel distances and pixel intensities, b_s^{opt} and b_t^{opt} are respectively the optimal shape and texture parameters over the training set. These parameters are used to describe the overall appearance variability of the modelled object, where *Ws* is a vector of weights used to compensate the differences in units between shape and texture parameters.

In the second approach, the combined model of appearance, subsequent to PCA over the appearance parameters, can be further written as:

$$c = \bar{c} + \varphi_{appearance} b_{appearance} \tag{3.6}$$

where $\bar{c}$ is the mean appearance vector, $\varphi_{appearance}$ is the matrix of appearance eigenvectors, and $b_{appearance}$ are the combined parameters.

Each parameter alters both the shape and texture aspects of the face. As in the shape PCA case or in the texture PCA case, we can further compress the combined model representation by removing the smallest eigenmodes. It is safe to consider

small-scale variation as noise [79]. It is common to conserve 95-98% of the allowed appearance variation. This rule was observed throughout our experiments.

In Figure 3.4 we present an example of a combination of appearance parameters that describes a new facial appearance. Like Figure 3.2 for the shape parameters we see that the final facial texture is built from a series of 2-D texture components analogous to the construction of a 1-D signal from a Fourier series.

Figure 3.4. An example of the combination of the 2-D AAMs appearance models, as shown in [78], that creates an unseen appearance of a face.

The main advantage of the combined model is that as we conserve only a percentage of the appearance variation, its number of resulting model parameters is always less than the total number of initial shape and texture parameters.

Historically, as the goal was to have the most compact/efficient model, the original work on AAM took this approach. However, in later work, researchers wanted to be able to separate out the shape and texture dependencies to better model specific facial variations, thus the independent form of AAM evolved later [79]. We will exemplify and explain later, in section 8.4, that if we train with a set of faces at different left/right poses, then the first shape parameters are mapped onto this variation as it is the most significant spatial variation. Later on, researchers showed that it is also possible to separate the main shape/texture subspaces into independent sub-subspaces [83-86].

In the next section, we describe the AAM search algorithm that enables us to find, analyse, and synthesise objects belonging to the same class as the statistical model of appearance.

3.3. The AAM Search Algorithm

After a Statistical Model of Appearance is created, the AAM algorithm can be employed to obtain an optimal fit of this model to a region of an image. Although the AAM algorithm can be applied to a full image we typically pre-process the image with a fast pre-filter to extract image regions with a higher probability to include an object which can be modelled by the statistical model (section 3.2). We remark that some applications of AAM have effectively used these models to perform face tracking without pre-filtering the input image. However there are faster algorithms to achieve an initial face detection and this is the approach we use.

Thus we assume that the AAM algorithm has been provided with a region of interest (ROI) with high probability to include a face. The search algorithm then determines parameters of the model which generate a synthetic image with an optimised similarity to the target image. This set of parameters defines the shape, position, and appearance of the target object in an image, and can be used for further processing, such as to make measurements or to classify the object.

During the optimisation stage, the parameters that we wish to determine are $p = \begin{bmatrix} g_s \\ b_c \end{bmatrix}$, where g_s are the shape 2D translation (t_x,t_y), rotation($\theta \in [0, 2\Pi]$), and isotropic scaling ($s>0$) parameters inside the image frame, and b_c are the combined model parameters.

By adjusting the AAM parameters p, the model g_m can be deformed to fit the image g_{img}, in the best possible way. The whole problem is treated as an optimisation problem in which we want to *minimise the difference between the new image g_{img} and the image synthesised by the appearance model g_m:*

$$\delta g = g_{img} - g_m \tag{3.7}$$

We evaluate the current error $E = |\delta g|^2$ and minimise it by varying the shape and texture parameters p with a predicted displacement. The quadratic error norm is applied as optimisation criterion:

$$E = \sum_{i=1}^{m} \left(g_m - g_{img} \right)^2 = \sum_{i=1}^{m} (\delta g_i)^2 = |\delta g|^2 \tag{3.8}$$

where m is the number of appearance parameters.

Although the model parameterisation can be compacted tremendously by the PCA (section 3.2), it is far from an easy task to optimise a system which is still of a high dimensionality. This is not only computationally cumbersome, but also theoretically challenging since it guarantees that the hyperspace sought is smooth [79].

To avoid computational complexity, a linear relationship between parameter changes δp and pixel differences δg is considered. Cootes et al. show that this simple approximation produces good results for the AAM algorithm [21, 87, 88].

This linear assumption is based on the following procedure. A first order Taylor expansion of the functions $\delta g(p)$ in the neighbourhood of the current point p is employed in order to linearise the problem:

$$\delta g_i(p+\delta p) = \delta g_i(p) + \delta p_1 \partial\left(\frac{\delta g_i}{\delta p_1}\right) + \delta p_2 \partial\left(\frac{\delta g_i}{\delta p_2}\right) \ldots + \delta p_m \partial\left(\frac{\delta g_i}{\delta p_m}\right) + O(|\delta g|^2) \quad (3.9)$$

Next, a rough approximation is made by completely neglecting the second order terms. So we obtain:

$$\delta g_i(p+\delta p) = \delta g_i(p) + R\delta p \quad (3.10)$$

where R represents the gradient of $\delta g_i(p)$ at point p.

The linear relationship between parameter changes δp and pixel differences δg can be assumed as,

$$\delta p = R\delta g \quad (3.11)$$

where R is a pre-computed ***matrix of prediction*** that is estimated during the model building stage. As this matrix is estimated only once, it makes the algorithm runtime efficient.

Estimation of the Prediction Matrix

In practice, R is estimated by a set of experiments on the training set. Each experiment consists of displacing a set of ground truth parameters by a known amount and measuring the difference between the model and the part of the image below the model. Then the image is resampled at the new prediction. A new error vector is

calculated and if the resulting error is less than the preceeding error, the new estimate is accepted. The procedure is iterative.

In his thesis, M.B. Stegmann [79] gave some practical values for the pose/model parameter displacements:

Pose parameter displacements

t_x, i.e. translation in the x direction: ± 6, ± 3, ± 1 (pixels)

t_y, i.e., translation in the y direction: ± 6, ± 3, ± 1 (pixels)

s, i.e., scaling: 95%, 97%, 99%, 101%, 103%, 105%

θ, i.e., rotation: ± 5, ± 3, ± 1 (degrees)

Model parameter displacements: $\pm$ 0.5, ± 0.25 for each parameter over the training set

Other values of displacements were proposed in the literature [82], but in our experiments, we employed Stegmann's update scheme with accurate results.

Figure 3.5 shows a practical example of the AAM search algorithm. The algorithm is applied for an eye model. The AAM eye model was implemented during our work and will be described in Chapter 5. Starting from a new, unseen picture for the training set, we want to find the eye parameters that best match it. Using the information given by the Viola-Jones face detector, an initial estimate for the eyes is given. Then, in an iterative process, we seek to minimise the difference between the new image and the synthesised one.

In practice, the algorithm required only two iterations. It can be noticed that in comparison with the original image, the eyes synthesised by AAM are blurrier because of the texture approximation (equation 3.11).

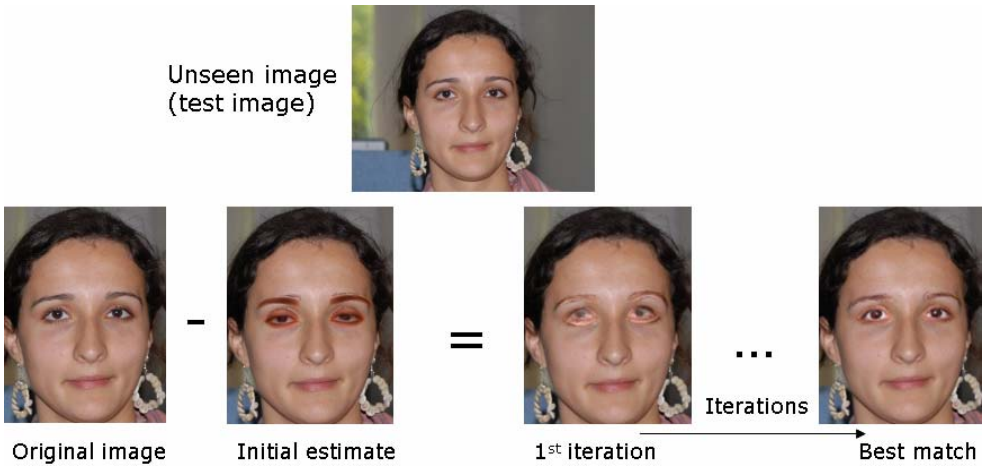


Figure 3.5. Example of the AAM search procedure to match an eye model on an unseen picture (see Chapter 5).

3.4. Challenges and Limitations of the Standard AAM

The main purpose for AAM is to find the appearance parameters in the case of an unseen image and to synthesise the object as close as possible to the target image. The AAM is a fast and accurate algorithm. Nevertheless, it presents several drawbacks that are going to be described in the current section.

Solutions to overcome these drawbacks are proposed in the literature. The most relevant references will be summarised in section 3.5.

Handling model generality

The AAM search algorithm is known to converge quickly and accurately if we have a picture similar to the ones in the training set [82]. Now as the size of the training set has to be restricted for practical reasons – remember the mark-up process for faces requires manual user tuning – there are limitations on the amount of facial variation we can incorporate into a single statistical model.

Thus difficulties will appear when we want to model faces which exhibit variations that are not represented in the training set. This is our biggest challenge – to model unseen images and unseen image conditions.

So the ultimate goal of face modelling is to develop a ***general*** Statistical Model of Appearance and an algorithm able to handle the ***generality*** of the test images. By a general model or generality of test images we mean no coarse constraints on the environment or on individual cooperation. For example we want to

be able to analyse human faces, despite race, age and gender differences, facial expression, pose, and various illumination conditions.

As you might expect such a goal is somewhat of a "grand challenge" and essentially unobtainable as there will always be extremes of facial appearance that are not incorporated into any finite sized training set. Nevertheless we can make some useful attempts to overcome some of the more challenging problems of face modelling such as handling generic lighting and pose variations for semi-frontal faces. Ultimately, as we will see in this thesis, we will need to apply more than one model – each model being specialised to a certain range of variations and even to modelling variations of specific regions or features within a facial region. In this way we can achieve useful progress in developing a more generalised solution.

Cluttered background

Another issue to consider is that AAM inherently models the foreground of the object and does not consider the background image. From their experiments, Sung et al. [19] concluded that AAM often fails to converge when the estimated object area in the target image includes the image background. This happens due to the misalignment that can occur when fitting an object. The inclusion of the background image makes the backward warped image inconsistent with the original appearance, which causes the algorithm to update parameters incorrectly and diverge.

This non-convergence of AAM becomes a significant problem when there are very bright backgrounds and/or textured backgrounds adjacent to the target object [89]. We remark that if pre-filtering techniques are applied to the ROI prior to applying the AAM it is generally not very expensive, computationally, to perform a foreground/background separation and to replace the background of the detected face region with a neutral background. In our research we have been careful to avoid test images with error-inducing backgrounds. We assume that in any practical implementation of our model that the appropriate pre-filtering would be employed prior to fitting the model.

Initialisation problem

Although the structure of a statistical model of appearance itself is simple, fitting an AAM to an image is a complex task that requires a nonlinear optimisation to find model parameters.

As discussed in section 2.4.1, AAM is very sensitive to initialisation. A faulty initialisation can bring errors that propagate, resulting in poor convergence, or even non-convergence of the algorithm. Typically, this initialisation is achieved by using a pre-processing method such as a global face detector, or a detector for salient feature points.

It can also be helpful to use a skin detector [90] to provide horizontal and vertical mean-axes for the face location. This will provide a sanity-check on the rectangular output from the face detector which may not always be accurately centred on the main face region. If this information is reliable, it can be added as an extra constraint to the search algorithm. However, there is the drawback of increasing the computational time. It would be helpful to find a compromise between the overhead of determining additional information about the ROI and the speed of the fitting algorithm.

Search algorithm limitations

Another limitation is the assumption of a ***constant*** linear relationship between the error image and the additive incremental updates of the parameters (see equation 3.11). As we might expect the assumption that there is such a simple relationship between the error image and the appropriate update to the model parameters is incorrect [91]. The result is that the existing AAM search algorithm often performs poorly in terms of the number of iterations required to converge and in terms of the fitting accuracy which can be obtained after convergence.

In summary, there are several limitations to both the model training stages and the AAM search process. Since Tim Cootes originally proposed the AAM modelling technique, there has been constant work aiming to extend its generality and to solve these limitations. Several of the most significant extensions proposed in the literature are summarised in the next section.

3.5. Extensions of the Standard AAM

In order to solve the challenges of a standard AAM, a large number of papers focus on new variations from its initial formulation. The variations refer to the building of

the statistical model of appearance, section 3.5.1, to the search algorithm, section 3.5.2, or to both stages.

3.5.1. Extensions of the Statistical Model of Appearance

AAM is mostly known in its 2-D version. When adapted for 3-D data, the approach is called the ***3-D AAM*** or ***Morphable Model (3DMM)*** [10, 92]. We mentioned this approach briefly in section 2.2.2.3. The search algorithm respects the standard AAM, customised for 3-D data. New face images can be registered by minimising the difference between the new face and its reconstruction by the face model function. The 3-D extension of the model is used as a means to overcome variations in *pose* and *illumination* [32].

Now, the 2-D AAM lacks the photorealism of 3DMM, but has two important advantages. It uses normal video rather than special-purpose scanning – which requires matching stereo images. Also, AAM can process video at full frame rates up to 30 fps, which makes real-time applications possible.

An alternative to the use of 3-D data is to use a set of models to represent appearance from different view-points, as the ***coupled-view AAM*** [34]. It has been shown [11] that to deal with full 180 degrees rotation, i.e., from left profile to right profile, one needs five models. The models are roughly centred on viewpoints at -90, -45, 0, 45, 90 degrees, where 0 corresponds to front-parallel. The pairs of models at 90 degrees, i.e., full-profile, and 45 degrees, i.e., half-profile, are simply reflections of each other. Thus only three distinct models are required. These models can be used for estimating head pose, for tracking faces through wide changes in orientation, and for synthesising new views of a subject given a single view.

Even when the pose problem is partially solved, fitting a face image for an individual who is not in the training set is often limited in accuracy, thereby restricting the range of application. This limitation mainly comes from the inability of AAM appearance counterpart to *generalise*, i.e. to accurately generate previously unseen visual data. One solution is based on ***re-training***, i.e., iteratively refitting the training data with the AAM and re-training, and is shown to improve the performances compared to the traditional single step AAM training [93].

Another solution is to combine a global model with a series of sub-models. These sub-models are typically component parts of the object to be modelled. This

approach benefits from both the generality of a global AAM model and the local optimisations provided by its sub-models. It is proposed in different versions, e.g. ***component-based AAM*** [94, 95], ***segmented AAM*** [96], or ***hierarchical decomposition*** [97]. In the current thesis, we pay special attention to this kind of representation. In our work we will demonstrate the benefits of applying such a component model are proved in the fields of facial expression decoding, pose representation, eye tracking, and eye modelling.

The statistical model of appearance has also been extended to exhibit temporal stationarity using conditionally linear processes for shape, motion, and appearance. This extension is beneficial when *tracking* a deformable object using AAM. In other words, rather than modelling only appearance, i.e., eigenfaces, only shape, i.e., ASM, or only motion, i.e., dynamic textures, we model all three characteristics simultaneously using the ***dynamic AAM*** [98] or the ***Pairwise Active Appearance Model (PAAM)*** [99].

3.5.2. Extensions of the AAM Search Algorithm

The AAM search algorithm has some limitations, as discussed in section 3.4. Many derivations of the algorithm have been proposed to increase its robustness to various factors, its accuracy, or its speed.

In one enhancement AAM and ASM are ***combined*** in a single search algorithm intended to combine the advantages of the two algorithms. This is claimed to result in a *more accurate fitting* than either of the two methods considered independently [100]. To integrate the local profile and global appearance constraints, the subspace reconstruction residual of the global texture is exploited to evaluate the fitting degree of the current model on unseen images. And, as in the standard AAM approach, global texture is used to predict and adjust the model parameters.

In order to solve the *generalisation* limitation, i.e., fit images which are not present in the training set, the ***Boosted Appearance Model (BAM)*** was introduced by the authors of [101]. Together with the conventional Point Distribution Model (PDM), a boosting classifier is used to iteratively update the shape parameters in PDM, based on a gradient ascent method, maximising the classification score of the shape-normalised image, i.e. of the image warped to the mean shape. A boosting-based classifier is trained so that is able to learn the decision boundary between two classes:

the warped images from the ground truth landmarks, and those from perturbed landmarks. The proposed framework is shown to improve the robustness, accuracy, and efficiency of the fitting procedure.

As mentioned in section 3.4, one of the challenges for AAM is a *cluttered background*. A robust AAM that uses active contour techniques [89] proved to be robust to a cluttered background. The proposed ***contour-based AAM*** search algorithm consists of two alternating procedures: (i) active contour fitting to find the contour sample that best fits the face image followed by (ii) the AAM search over the best selected facial boundary contour.

The original definition of AAM deals only with the segmentation of the input image at the original resolution level. A ***hierarchical***, or ***multi-resolution***, processing is proposed by the authors of [102] to deal with situations when the optimisation process becomes trapped in local minima. For that, a multi-resolution scheme using a Gaussian image pyramid is employed to optimise the model parameters in a coarse-to-fine manner. By applying this approach, *computational complexity* can be reduced and the quality of the results improved.

In order to increase the algorithm *speed*, we can reduce the number of parameters that need to be optimised. ***Direct Appearance Model (DAM)*** [103] uses the texture information directly to predict the shape and to update the estimates of position and appearance. In this way, DAM includes admissible appearances previously unseen by AAM and improves the *accuracy*. It is in contrast to the AAM idea of modelling the appearance subspace from shape and texture combined. DAM cuts down the memory requirement to a large extent, and further improves speed.

The ***shape-driven AAM*** [88] uses the image residuals for driving only the pose and shape parameters, while the texture parameters are directly estimated by fitting to the current texture. The convergence *speed* is increased by reducing the number of parameters that need to be optimised.

Another limitation of AAM is the convergence scheme which assumes that the gradient matrix is fixed around the optimal coefficients for all images (equation 3.11). Such a fixed gradient matrix inevitably specialises to a certain region in the texture space and is no longer a good estimate of the actual gradient as the target texture moves away from this region. Hence, an ***adaptive AAM algorithm*** [104] that linearly adapts the gradient matrix according to the composition of the target image texture is

proposed. This adaptation search for *a better estimate for the actual gradient* and is performed by adding linear modes of change to the gradient matrix.

The original formulation of the AAM is additive from the point of view of parameters update, but makes the assumption that there is an approximately constant relationship between the error image and the parameter updates. This assumption is not always satisfactory. The ***Inverse Compositional*** [105] update scheme leads to an algorithm which requires only linear updates, so it is as efficient as the AAM, but which is a true gradient descent algorithm. The initial, general problem is to optimise the expression in (3.12) with respect to the parameters p:

$$\sum_{x} [A_0(x) - I(W(x,p))]^2 \qquad (3.12)$$

There are more versions of image Alignment, such as the Lucas-Kanade method, the Forward-Compositional approach, and the Inverse-Compositional approach.

3.5.3. Discussion

Extensions to the appearance model and to the AAM search algorithm span a large range of applications and purposes. In section 3.5, we have described some of the most significant AAM extensions.

Every extension from the basic approach has its advantages and drawbacks and it was created in order to compensate for a specific challenge. A direct comparison between AAM extensions would be interesting and helpful. It is not an easy task to directly compare results of the AAM extensions, due to different training sets used in the model creation and to different test sets used for the model assessment. In fact there are not many comparisons in literature between AAM extensions.

In [88], Cootes et al. compare the sub-sampling and the shape-driven AAM with the original formulation. Both extensions are aimed at improving the speed and robustness of the AAM search. From their experiments, they determined that both extensions lead to faster convergence, but are also more prone to failing to converge. The shape based method is able to locate the points slightly more accurately than the original formulation. The authors conclude that it may be possible to use

combinations of these approaches to achieve good results quickly, for instance using the shape based search in the early stages, then polishing with the original.

The way in which AAM is implemented, either in its original formulation or in one of its extensions, depends on the application requirements. In our work, we developed AAM models for various consumer applications, e.g., facial expression recogniser, blink detector, smile detector. These types of applications require a precise modelling of facial features.

The component-based AAM [94] proved flexible, in order to adapt to all possible types of facial variability typical to facial expressions (see Chapter 8 for a more detailed discussion). In our work we have adopted this approach, modifying and extending it to accommodate our applications.

In our work, we also address the challenge of pose variation, inherent in consumer pictures illustrating real-life conditions. In order to deal with this challenge we propose extensions of the standard AAM [83]. These extensions will be detailed in Chapter 8.

In the next chapter we are going to review published research that investigates human facial expression modelling, classification, and recognition, based on AAM.

Chapter 4 - Literature Survey: Facial Expression Modelling using AAM

The purpose of the current chapter is to review published work that uses AAM in expression modelling, classification, and recognition. Our choice was motivated by the popularity and the efficiency of the method in the facial expression modelling field and by our earlier experiences in using the AAM approach to model faces and facial characteristics.

We start the chapter by analysing emotion recognition in the field of psychology and neurosciences (section 4.1). We search to understand to what extent is expression recognition influenced by gender, familiarity, symmetry, or expression intensity. This study helps us to better understand facial expression analysis and to identify its challenges. Ultimately the most useful outcome of such an analysis is to provide an accurate determination of the emotional state of the person under observation. Thus we pay particular attention to research which relates facial expressions to the underlying emotional states of a person.

Our work is centred on expression analysis using AAM. In section 4.2, we survey the applicability of the method in the expression analysis field. Different AAM extensions proposed in the literature, for still images or video sequences, are summarised. We searched to emphasise the novelty of certain techniques proposed in our work, by comparing it with the existing literature. This section ends with a qualitative comparison of the results obtained in literature. In section 4.3 we explain expression classification based on AAM and compare results for different types of classifiers. Expression recognition is explained in section 4.4.

Details of several research databases designed to provide facial images exhibiting a comprehensive range of expressions are presented in section 4.5. These databases are widely used for testing automated expression analysis algorithms and systems and provide a useful reference framework for the experiments documented later in this thesis. We remark that existing databases were not entirely satisfactory for our research as a significant aspect of our research was focussed on providing a solution which is sufficiently generic to apply to unconstrained images. Thus we had to develop our own database to provide a more challenging test baseline for some of our experiments.

4.1. Psychological Research on Emotion Recognition

Facial expression analysis is a multidisciplinary concept. To be able to develop an automatic system recognising human expressions, one should first understand the concept from social, neurological, and psychological points of view. Researching these fields, we studied several factors that influence facial expressions that we describe in this section. To what extent is the recognition of expressions influenced by factors such as, gender, age, race, facial familiarity and symmetry, type and intensity of emotions? In the current section, we try to answer this question and to explain how these factors influence our work hypothesis.

Gender

Studies in psychophysics have shown that there is a relation between facial expression and gender. Shrum [106] discovered that women generally perform better at expression recognition. He states that the difference in the amount and intensity of one's facial expressions is also gender related, with women more frequently and more intensely expressing certain emotions than men. This might be also explained through cultural influences. In modern Western society women are expected to be more emotionally extrovert, whilst men are required to maintain a more reserved face [107].

In our work we want to obtain robust expression recognition for both genders. However, we will investigate gender-specific differences in the appearance of facial expressions (see section 9.3.2).

Age

Age influences facial appearance and in consequence, face expressions are also affected. For example, infants have smoother and less textured skin [108]. The reduced texture of infants' skin, their increased fatty tissue, juvenile facial structure, and lack of transient furrows may contribute to the differences observed in expression analysis between infants and adults.

In our work, we aim at accurate expression decoding rates for subjects across a range of ages.

Ethnicity and culture

Research has uncovered that culture is a strong determining factor when interpreting facial emotions. A study performed by Yuki et al. [109] reveals that in cultures like the Japanese one, where emotional control is the standard, eyes are considered the main source of interpreting emotions. Investigating this phenomenon in the United States [109], a culture where emotion is much more openly expressed, it was determined that the focus is on the mouth.

Elfenbein et al. [110] showed that cultural familiarity is associated with greater accuracy in emotion recognition. Individuals can more effectively understand universal emotions expressed by members of a cultural group to which they have greater exposure. However, the universal emotions defined by Ekman and Friesen [42] (see section 2.3.1) are considered to be universal across human ethnicities and cultures.

From the computer vision point of view, ethnicity and cultural diversity is present by the fact that some cultures might involve wearing beards, eyeglasses, or jewellery, considered as sources of facial occlusion.

Symmetry and asymmetry of the face

Facial asymmetry is considered an important factor in evaluating facial and expression recognition [111]. Sackeim conducted the following study [111]: pictures of human faces posing different emotions and their mirror reversals are split down the midlines, and left-side and right-side composites are constructed. Subjects judged left-side composites as expressing emotions more intensely than right-side composites. The findings indicate hemispheric asymmetry in the control over emotional expression in the face.

Mendolia [112] underlines that negative emotions are more easily perceivable from the left hemiface than from the right hemiface, and that, in general, the left hemiface is perceived to display more emotion than the right hemiface.

In our work, we did not explicitly consider the asymmetry contribution to facial expressions.

Expression intensity

Expressions must be seen as dynamic actions. They occur at different levels of intensity and at different moments in time. A study conducted by Hess, Blairy, and

Kleck [113] involved the measurement of the influence of the intensity of facial display on the accuracy with which the facial expressions are decoded. The results revealed that the perceived intensity is directly correlated with the pictured expression intensity.

In our work, we present some preliminary results in decoding expression intensities, described in section 9.2. We simplify our approach, by considering only two intensity levels for each investigated expression. Note that most of our work is focussed on static expressions. This may be considered a weakness of our techniques, but there are many applications where live images or video is not available, yet it is still desirable to interpret a person's facial expression and attempt to determine their corresponding emotional state. And, arguably if we can achieve high recognition rates from still images it will be easy to further enhance recognition from dynamic image sequences.

Expression categories

Not all facial expressions are used with the same frequency and some are more difficult to perform [114, 115]. Tamminen et al observe that, on average, the easiest expressions to perform, e.g., happiness, are the most recognisable. Also, negative expressions, such as sadness and fear, do not achieve the same levels of intensity as the positive ones, such as happiness, and as a consequence, they are not as distinctive [115].

In our experiments, summarised in Chapter 8, we also found happiness to consistently be the most recognisable facial expression. Thus, these studies in [114, 115] corroborate our results.

4.2. AAM and its Applicability in Expression Modelling

In this section we describe the use of AAM to model facial expressions.

Firstly, by reviewing the literature, we evaluate the standard methods of applying AAM when modelling facial expressions (section 4.2.1). We observe and discuss inherent challenges of this standard approach.

In section 4.2.2, we survey proposed solutions in the literature. There are different types of proposed AAM extensions that aim to accurately model facial expressions. Some researchers separated the sources of facial variation, such as face

expression. This variant is presented in section 4.2.2.1. Then, in order to obtain a more precise detection of individual facial features, other researchers have made use of facial sub-models. When dealing with facial expressions, local features are responsible for most of facial variation, as one might expect and as we also found in our experiments, described in detail in section 8.6.1. Section 4.2.2.2 summarises approaches that use face sub-models, rather than a global face model.

Alternatives to the standard shape and texture AAM formulation were also proposed (section 4.2.2.3). Section 4.2.2.4 presents attempts to combine AAM and ASM to take advantages of the benefits of each model, while in section 4.2.2.5 we present 3-D AAMs applied to model expressions.

In addition we summarise research on AAM extensions that track facial expressions (section 4.2.3).

In section 4.2.4 we discuss all of these AAM extensions or enhancements for modelling facial expressions and determine how best to build on this state of art research and know-how. This will lead to our own research contribution described in Chapters 8 & 9.

4.2.1. Modelling Face Expression using the Standard AAM Formulation

In section 2.3.4 we concluded that when using AAM, facial features are detected quickly and accurately. In addition we concluded that this technique has a potential value in facial expression applications.

Various research work presented in the literature explores this topic of using AAM in the area of expression modelling. When using the standard AAM formulation, a first observation is that the model does not separate facial expression from other sources of model variability. This fact affects the facial expression decoding performance [84, 116]. This conclusion is further supported by the results obtained in [84] and summarised in Figure 4.1.

	AAM + LDA (first 20 modes)	Bilinear asymmetric
neutral	55%	77.50%
anger	66.67%	77.78%
disgust	57.14%	71.43%
fear	60%	80%
joy	68.42%	94.74%
surprise	92.86%	92.86%
sadness	100%	88.89%
total	67.59%	83.33%

Figure 4.1. Facial expression recognition rates obtained in [84], when using the standard AAM formulation (first column) and when using extension of the algorithm (second row).

From our experiments (see section 8.6.2) we determined that in the standard AAM formulation, where we deal with a holistic model, faces can be too closely constrained to the principle model variations, such as pose and illumination, learned during the training phase. From a practical perspective this causes difficulties when we deal with facial expressions, where more subtle local features are responsible for the facial variation that we wish to discriminate. This suggests that alternatives to the standard global representation of a face should be found.

As already discussed in section 3.4, studies show that AAM is sensitive to pose and illumination variations [32]. Extensions of the model that solved the pose and illumination challenges also obtain more accurate results in expression modelling [117, 118].

4.2.2. Modelling Face Expression using Extensions of the Standard AAM

As a consequence of the inherent challenges of the standard global AAM formulation, many alternatives were proposed in literature. In the following sub-section we are going to summarise the extensions from the literature that we consider as the most representative. For each we summarise the core method and comment on how it improves the accuracy of the model fitting and also the principle drawbacks.

4.2.2.1. The Bilinear AAM and Related Methods

Initial estimates of the subspaces defining the sources of variability, i.e., lighting, pose, identity, and expression, are obtained in [119] by applying PCA on groups of images displaying those sources of variability. Four sets of images, each showing major variation on one subspace alone were used for training. The authors then applied an expectation-maximisation algorithm to maximise the probability of coding across these non-orthogonal subspaces.

Aboud and Davoine [84, 85] use a global AAM in conjunction with two bilinear factorisation models to separate expression and identity factors from the global appearance parameters. The results obtained in [84, 85] are highly accurate. However, in their approach it is first assumed that identity corresponds to the first training identity. This assumption limits the model capability to generalise for new subjects.

In [86], Macedo et al. separate expression variation using multilinear analysis, together with AAM. The authors state that by applying a multilinear analysis procedure over shape and texture AAM parameters, they are able to separate the identity and expression factors hidden in these parameters. The authors use this approach to develop a facial expression transfer application, i.e., mapping a facial expression performed by a given subject in a picture, onto the photograph of another person's face. Figure 4.2 explains this application and exemplifies some results.

Figure 4.2. Expression transfer between photographs, as realised in [86]. The first picture provides the expression that will be transferred in the second picture. Results in the Lab and, respectively, RGB colour spaces are shown in the last two pictures.

Alternative methods of deriving quasi-linear models are also able to successfully retrieve expression parameters. For example, Buenaposada et al. [120] introduce a subspace representation of face appearance which separates expressions from illumination variations. Then, the appearance of the face is represented by the addition of the independent linear subspace modelling expressions and the one modelling illumination. It requires two image sequences; in one sequence a certain

expression is subject to all possible illuminations, while in the second image a face under one illumination performs all facial expressions.

Mercier and Dalle [121] separate sources of variability by introducing a model that models the face as a sum of an identity part and weighted expression parts. It can be used both for applications in facial recognition and conversely as a face anonymiser by removing the identity component from a facial image while retaining the facial expression, as displayed in Figure 4.3.

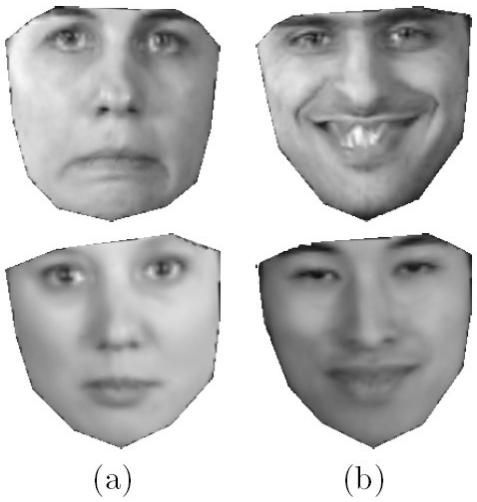


Figure 4.3. The results obtained by the face anonymiser, as illustrated in [121], with the model when it fails (a) and when it succeeds (b). The first row contains the faces to anonymise, while the results with the modified identity part are on the second row.

4.2.2.2. Decomposition of the Global AAM into Sub-models

Component-based AAM [94] has been employed as a solution to the drawback of using a holistic AAM model. It combines a global model with a series of sub-models. These sub-models are typically component parts of the object to be modelled. This approach benefits from both the generality of a global AAM model and the local optimisations provided by its sub-models.

Zalewski and Gong [97] define expressions as a combination of intrinsic functionalities of subcomponents. They provide a hierarchical decomposition as follows: jaw outline, nose and centres of the eyes, and the mouth form the root of the hierarchy; as branches of the decomposition they use eye-eyebrow pairs and mouth. The root component is used for estimating pose. The branches are used for expression modelling.

In [122], a component shape model, such as mouth shape and eye-contour shape, is used in addition to the global shape model to achieve a stronger representation for facial expression components, when subject to complex pose variations which tend to dominate the model.

The idea of a hierarchical model is also employed in [123], so that major appearance modes of each sub-facial model encode the major variation for the corresponding area. For example, the highest mode of variation in the left-eye model encodes a blink.

In our work, and inspired by some of this research from the literature, we adopted the idea of decomposing the face model into several sub-models corresponding to relevant facial features. A detailed description of our method, its motivations and its enhancements over techniques described in the literature will be presented in section 8.6.2.

4.2.2.3. Alternatives to the AAM Shape and Texture Representations

In order to improve the expression recognition rate, Li et al. propose [124] alternatives to shape and texture information. Face texture, intrinsic geometry, and expression information are modelled separately, by fitting a generic mask (see Figure 4.4). Subsequently, texture and geometry attributes are re-combined to form a classifier capable of recognising faces with different expressions.

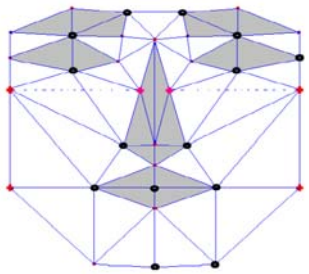
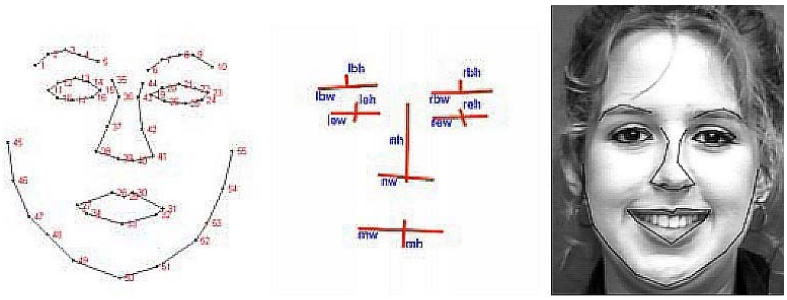


Figure 4.4. The face masked proposed by [124].

Another AAM extension, intended to improve expression modelling, is presented in [84, 85]. Shape vectors are multiplied by appropriate weights to compensate for differences between pixel distances and intensity values. Linear Discriminant Analysis (LDA) is applied to extract the six most discriminating features which maximise class separability.

Discrete Choice Models (DCM) [125] is a concept known from econometrics. In this paper the authors redefine facial expression classification as a discrete choice

process. Facial expressions are described with expression descriptive units (see Figure 4.5), consisting of a set of high level features derived from an AAM. A new classifier based on DCM and a set of expression descriptive units of the facial components, i.e., eyebrows, eyes, nose and mouth, and the mutual interactions between them, are also proposed by these authors.



(a) (b) (c)

Figure 4.5. The description of the DCM as shown in [125]: a) Facial landmarks (55 points);b) Facial components descriptors; c)Expressions Descriptive Units.

4.2.2.4. Combining AAM and ASM and Other Extensions

ASM together with AAM has been widely recognised as a valuable method for image understanding [20, 88]. ASM was described in detail in Section 2.2.2.2.

One advantage of ASM is the fact that shape parameters bring important information in the case of non-rigid facial variations [97, 115]. Shan et al. [59] enhance the ASM algorithm with global texture constraints. This approach combines the ASM local profile matching with the AAM global texture model, taking advantage of both techniques. Each landmark is firstly matched by its local profile in its current neighbourhood, and the overall configuration of all the landmarks is re-shaped by the statistical shape constraints. Then, the global texture is warped out from the original image. The texture information is used to predict and update the shape parameters before executing the next local match iteration for each landmark point.

Both ASM and AAM use PCA to model texture. One drawback of PCA is that it does not consider discriminative criteria between samples – feature points, and negative samples – non-feature points [58]. As a consequence, its local searching stage often falls into local minima, thus failing to converge correctly.

An alternative is presented in [58], where classifiers using look-up-table (LUT) Haar-like features are trained from a large data set to model texture. The LUT Haar-like features are proposed to compensate for this drawback of the PCA

technique. These features are constructed based on 2-D Haar-like features, corresponding to simple rectangle patches and they can be calculated efficiently using an integral image representation [126] of the original image. The generality and robustness of the boosting method guarantee the performance, when dealing with expression changes.

4.2.2.5. 3-D AAM Deformable Model

AAM is fundamentally a 2-D affine model. Several 3-D versions were proposed in literature [10, 92, 127, 128], as a means to overcome variations in pose and illumination [32] and to describe facial expressions more into detail.

The debate between modelling faces using 2-D AAM or 3-D AAM remains open. Recognition systems using 2-D AAM have been able to claim early success especially for constrained types of facial dynamics. Some of the challenges for 2-D AAM image segmentation and recognition are viewpoint changes or self-occlusions. These challenges come from the fact that less a-priori knowledge about the kinematics and shape properties of the faces are known.

Usually, in 3-D AAM, researchers use data sets of faces acquired with a 3-D scanner [10]. 3-D-Anthropometric-Muscle-Based AAM (AMB-AAM) [92], Candide 3-D AAM [127], or cylindrical representation of the face [128] are just several ways to include 3-D animation in a standard 2-D AAM.

Classical AAM parameterises facial shape in 2-D defined by labelled landmark values. AMB-AAM [92] employs a different approach for 3-D data: the shape is linearly controlled by anthropometric parameters. These are parameters dedicated to the measurement of the size and proportions of human faces that control facial expressions and facial types. The Anthropometric Action Units (AAUs) used are: facial width, mandible width, chin height, nose vertical shift, nose height, nose width, mouth vertical shift, mouth width, upper lip height, lower lip height, and eyebrow vertical shift. The AMB-AAM trained model appears more compact than AAM and it presents less shape parameters. Results show that the AMB-AAM fitting results are more accurate than those of AAM on 78% of cases [92].

In [128], the head is modelled as a texture mapped cylinder (see Figure 4.6). The authors compared the effectiveness of their approach as opposed to a standard planar model. Their cylindrical model out performed the planar model in all the

experiments in terms of precision and stability. It is also not as sensitive to errors in the initial positioning of the model as the planar model.

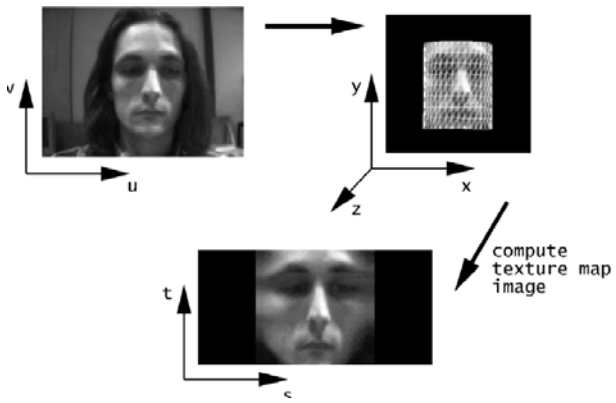


Figure 4.6. Mapping from image plane to texture map, as proposed in [128].

The Candide 3-D model is also proposed in the field of face expression, due to its simplicity [129]. In this representation, shape is fully described by the concatenation of the 3-D coordinates of face vertices. The model is then defined as the linear combination between Shape and Animation Units. Shape Units provide a way to deform the 3-D mesh such as to make the head or the eyes wider, etc. Animation Units provide a way to deform the 3-D mesh according to some predefined facial animations.

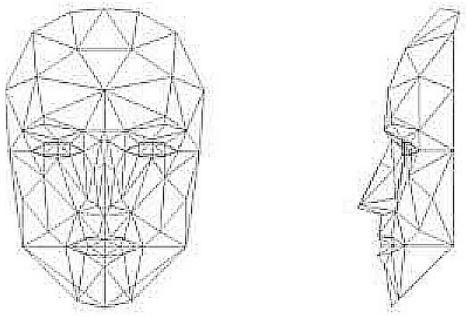


Figure 4.7. Candide-3D with 113 vertices and 183 triangles, as described in [129].

4.2.3. Tracking Face Expression using AAM Extensions

Facial expression analysis can be performed from static images [84, 85, 114, 116, 119-121, 124, 125, 130-134] or from video sequences [135-142]. An expression tracking system estimates the rigid and the non-rigid motion, e.g., facial expression, of an object through an image sequence [136]. The identity factor remains constant, while illumination, pose, and expression parameters vary, each with its own dynamics. So, in case of tracking, a person-specific model is more appropriate [143].

The output of the model from previous frame is usually used as an initial approximation of current parameters.

There are several methods described in the literature which adapt AAM techniques for expression tracking. Methods such as the Kalman filter [139], a low dimensional manifold embedding separated expression parameters [142], Resolution-Aware Fitting [138], or image differences, based on Discriminant Non-negative Matrix Factorisation [140, 141], have been used.

One successful approach to catch the dynamics in faces, in terms of facial expressions, is the Candide model [135, 136] (also see section 4.2.2.5). The tracking algorithm consists of decoupling rigid motion, i.e., head pose, from non-rigid motion, i.e., facial expression. After estimating head pose, an AAM search is used to determine and track the vector of animation parameters [144].

In an alternative approach to simplify the complexity of global parameters, Chen and Davoine [137] analyse local regions of several representative features. During tracking, feature-based local distributions are obtained directly from the video frames by comparing the local distribution with a pre-computed model.

A modular eigenspaces is developed in [143] bringing benefits over global eigenspace methods, e.g., more accurate reconstruction of the regions of interest, lower computational costs, and robustness to occlusions. The authors use predefined masks, i.e., appearance models that decompose face images within regions corresponding to the eyes, mouth, and remainder of the face (see Figure 4.8). The appearance within regions varies independently. In the current implementation the regions move together according to a single affine model.

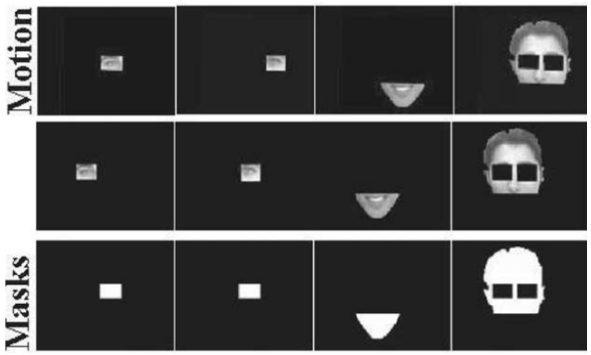


Figure 4.8. In [143] face images are decomposed using appearance models within regions corresponding to the eyes, mouth, and remainder of the face.

Choi and Oh [145] improve the AAM fitting algorithm employing a second order minimisation approach. The authors prove that this approach gives the ability of

correct convergence with a slight loss of frame rate to AAM. The correctly extracted facial shape with AAM ensures that facial expressions do not experience large errors.

4.2.4. Discussion of Prior Research from the Literature

Following our literature review on facial expression analysis there are a number of key questions we need to answer: (i) Which AAM parameters are more relevant to expression decoding? (ii) Is a local approach more efficient in modelling expression than a global approach? And finally, (iii) what are the future directions in facial expression analysis?

Let us begin with the first of these questions:

(i) Which AAM parameters are more relevant to expression decoding?

AAM extracts shape, texture, and the combination of the two, i.e., appearance parameters. It is still debatable which of the AAM parameters carries most of the expression information. Researchers agree that shape features have a large role to play in facial expression recognition [68, 97]. The question is whether texture brings sufficient new information in order to justify the additional parameters that it adds to the model.

In the literature, opinions are divided. In [97], after experimenting with shape, texture and AAM parameters, the shape component is preferred. The explanation is that texture information is more susceptible to external factors such as illuminations or identity changes.

Kotsia et al. [140] state that when using appearance parameters, the results are significantly better than when using either texture or shape information. The authors motivate the affirmation by the fact that the introduction of texture eliminates some of the confusions observed when using shape information only. It is stated that shape information is not enough to fully describe expression effects, such as furrows and wrinkles. The authors conclude that texture can capture all the necessary information where the shape description would fail, thus making the fusion of the two types of information more powerful.

In section 8.3 we will describe experiments that we performed concerning this topic.

(ii) Is a local approach more efficient in modelling expression than a global approach?

Another disputable aspect is the use of global facial appearance versus local features. Holistic approaches are preferred to local approaches when modelling faces [19-23].

However, facial expressions are defined by the dynamics of the individual facial features, as we experimented in section 8.6.1. These features have to be accurately modelled in order to capture facial expressions. Sub-models modelling these features along with the global model can help to improve the performance of an expression recognition system [94, 97, 122, 123]. This type of representation [38] also promises to solve pose issues.

In our work, we ultimately propose such a combination of local sub-models with a holistic, global model and demonstrate its benefits over other techniques (see section 8.6.2).

(iii) Future work directions?

In conclusion, there are still unresolved disadvantages to using AAM in expression modelling. Citing other research articles, we mention poor quality of the reconstructed images, manifesting as blur in the eyes and in the mouth area [146], poorly modelled features, e.g., wrinkles, tattoos, piercings, and birthmarks [56], low accuracy against high variations caused by unknown occlusions and illumination changes [131].

4.3. Facial Expression Classification

In Chapter 2 we overviewed an automatic facial expression analysis system (see Figure 2.1). The last module of such a system typically performs expression classification and recognition.

Expression classification is achieved by a classifier that uses measurable feature vectors to assign an object to one of the two corresponding classes. These feature vectors are extracted from input data. In our case we use AAM techniques to extract the relevant facial features – represented as parameter values of the AAM model. Data samples are then classified based on these parameter values.

4.3.1. Classification Challenges

There are several aspects that have to be considered when using a classifier. Its training is one of the most important steps. A careful choice of the type and number of features used to train the classifier is important. On one hand, when the training set is too small, the classifier does not learn enough useful information about the different classes from the particular training set. This effect is also known as ***under-training***.

On the other hand, when using a very large number of training examples, many classifiers may obtain good performances on this training set, but perform poorly on examples outside it. This is also known as ***over-training***. The classifier is not able to generalise well, becoming too specific to a particular set of training data. A very large number of training examples also make classifiers computationally expensive. A compromise between sufficient training data and avoiding a very large number of training examples has to be found, often by trial and error.

Practical examples of these challenges will be given in Chapter 8.

4.3.2. Examples of Classifiers used in Expression Classification

In the literature, most facial expression classification is reduced to a classification between the six universal expressions and the neutral one (details are given in section 2.3.1). We chose to adopt this approach in our work.

Different types of classifiers are employed in the literature. Artificial Neural Networks (ANN) [56, 147], Enhanced Fisher Classification Models (EFM) [57, 148, 149], Support Vector Machines (SVM) [107, 115, 118, 125, 150], Nearest Neighbour (NN) based on Euclidian [151, 152], Mahalanobis [75, 153], and cosine distances [151], Hidden Markov Models (HMM) [154], Linear Discriminant Analysis (LDA) [125], Bayesian classifications [155], and several other methods have been successfully used. We give some relevant examples, mentioning also published work that describes their implementation.

In our work, two classifiers are compared: NN (section 8.4.1.1) and SVM (section 8.4.1.2). The choice of the two classifiers is based on their positive results obtained in the literature [156].

4.3.2.1. Nearest Neighbour (NN)

One of the simplest classification schemes, that proved though efficient in expression classification is the NN classifier based on distance similarity. The theoretical description of NN is based on [157].

NN are well-suited for applications where there are many classes and a reduced number of examples per class, as in the case of facial expression recognition. NN are also suited when we have no apriori knowledge of the distribution of observations and classes other than that provided by pre-classified samples. In this case, a decision to classify a sample into a specific category will depend only on a collection of correctly classified samples.

NN are basic classifiers which do not require a training process. The NN rule classifies a sample a in the category of its nearest neighbour. More precisely $a'_n \in \{a_1, a_2, ..., a_n\}$ is a NN to a if $min\ d(a_i,a)=d(a'_n,a)$ where $i=1,2,...,n$. If the number of pre-classified points is large it makes good sense to use the majority vote of the nearest k neighbours instead of the single NN. This method is referred to as the k-NN rule. A common choice for the metric is the Euclidean distance. Other alternatives include cosine and Mahalanobis distances.

The ***Euclidian distance*** between to p-dimensional vectors $x=[x_1, x_2, .., x_p]$ and $y=[y_1, y_2, .., y_p]$ is defined as:

$$D(x,y) = x - y = ((x_1 - y_1)^2 + ((x_2 - y_2)^2 + ... + (x_p - y_p)^2)^{0.5} \qquad (4.1),$$

while the ***cosine***, or ***covariance distance*** is defined as the cosine of the angle between two vectors:

$$d(x_i, x_j) = \cos\Theta = \frac{x_i^T x_j}{\|x_i\|\|x_j\|} = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\left(\sum_{k=1}^{m} x_{ik}^2\right)^{1/2} \left(\sum_{k=1}^{m} x_{jk}^2\right)^{1/2}} \qquad (4.2)$$

where Θ represents the angle between the two vectors. It equals 1 when the vectors are identical, and it is 0 when the two vectors are orthogonal.

The cosine distance can be also defined as the angle Θ between the two vectors, i.e., $d'(x_i,x_j)=arcos\ d(x_i,x_j) = 1-d(x_i,x_j)$. The cosine distance ignores absolute sizes of the measurements and it only considers their relative sizes.

Different types of distance metrics were used in expression analysis, such as the Euclidean distance [151, 152], the cosine [151], or the Mahalanobis distance [75, 153].

Bartlett et al. [151] compared cosine and Euclidian distances for different architectures. The authors prove that ICA performs significantly better using cosines distance rather than Euclidean distance, whereas PCA performs the same for both distances.

In our work, we also evaluated this type of classifier (see Chapter 8).

4.3.2.2. Support Vector Machines (SVM)

These appear to be one of the most popular forms of classifier when dealing with facial expression. SVMs, introduced by Boser, Guyon, and Vapnik in 1992 [158], have been accepted as excellent classifiers for a wide variety of problems. By effectively embedding vectors in a higher dimensional space, SVMs can solve highly non-linear problems.

A detailed description of the SVM techniques can be found in [159]. SVM is a widely used classification method that presents several important advantages. SVM allows many features even if the training material contains few examples. SVM classifiers can be trained to maximally discriminate a particular class. All these advantages recommend SVM for expression classification/ recognition and in consequence it is a widely used classifier in these tasks.

SVMs are more complex than NNs and involve a training stage. They operate by finding a hypersurface in the space of possible inputs. This hyperplane attempts to split the positive examples from the negative ones. The choice of the hyperplane is based on the maximum distance to the nearest of the positive and negative examples. When classes cannot be linearly separated in the original input space, the input space is nonlinearly transformed into a higher dimension feature space. This procedure is done by a nonlinear kernel function, e.g., polynomial, Radial Basis function (RBF), multi-linear perception. Then, the linear optimal must be found to separate the hyperplanes.

Various methods for hyperplane separation can be used, i.e., Quadratic Programming, Sequential Minimal Optimisation, and Least-Squares.

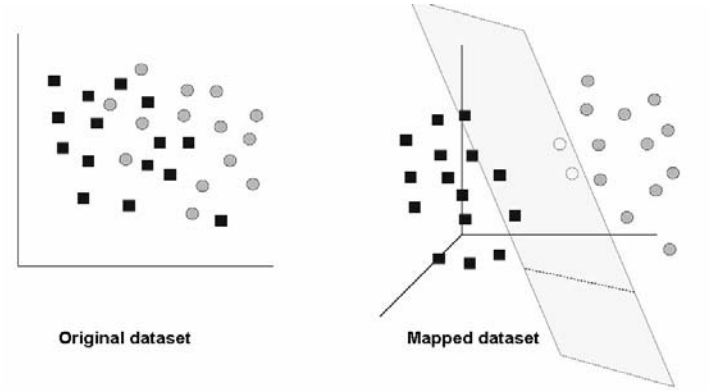


Figure 4.9. SVM algorithm.

Beszedes and Culverhouse [115] adopt SVM for AAM vectors classification and they use the human visual system (HVS) as ground truth for comparison. The decrease of image sizes reduces the classification accuracy for both SVM and HVS. The SVM classifier outperforms the average classification accuracy of human respondents in case of expression classification for all the tested image sizes. However, the SVM fails in cases of expression level intensity classification.

In their research, Bartlett et al. [118] apply machine learning methods such as SVM and AdaBoost, to texture-based image representations. The output margin for the learned classifiers predicts action units (AU) intensities. Frame-by-frame intensity measurements indicate facial expression dynamics that were previously intractable by human coding. For their system, AdaBoost performed slightly better than SVM.

More details on SVM and experiments based on its implementation will be given in Chapter 8, as SVM represents our choice for classifying facial expressions.

4.3.2.3. Discussion

The literature lacks any comprehensive comparisons between different methods of expression classification.

According to Wilhelm et al. [160], who compared results between NN, Multilayer perceptron (MLP), Radial Basis function (RBF), and Learning Vector Quantisation (LVQ) networks, the best results were obtained with AAM and MLP classifiers or ICA with NN classifiers, respectively. Another comparison [125] affirms the superiority of Discrete Choice Models (DCM) 78.26% and SVM 76.52%, against General Discriminant Analysis (GDA) 62.60% and against Linear Discriminant Analysis (LDA) 49.56%.

In Table 4.1, we summarise the accuracies for various classification methods. We only compare results for classifiers using AAM features as input. In this way, we only compare performances depending on the accuracies of the classifiers and the methods of training/ testing the classifiers.

Perhaps the most popular classifier in expression classification is SVM. A strong motivation of using it in expression classification is its high classification performances and its consistency. In our work, we also use this classifier. For comparison, we also implement NN based on Euclidean and cosine distances. Results of both classifiers will be detailed in Chapter 8.

TABLE 4.1. ACCURACY OF VARIOUS METHODS AND THE CONDITIONS OF TESTING.

System	Classifier	Acc(%)	Database	Conditions of testing
[161]	ANN	89%	Karolinska Directed Emotional Faces	-training: 1512 appearance vectors -3-layer feed-forward neural network, with 94 input neurons, 15 hidden neurons and 7 output neurons
[147]	MLP	99%	Their own video sequences	- person-specific oriented - 7 input layers, 5 hidden layers, 5 output layers
[56]	ANN	89%	Karolinska Directed Emotional Faces	- training: 980 facial images
[148]	EFM	87%	Cohn-Kanade	- 5 Cross Validation method - 4 expressions tested
[57]	EFM	88.9%	Cohn-Kanade	- build AAM on 272 facial images, AAM fit on the remaining 172 images
[149]	EFM	87.2%	Cohn-Kanade	- Inverse Compositional Image Alignment method - 4 expressions tested - AAM: 498 face images; test: 5 Cross-Validation method
[150]	SVM	90%	Cohn-Kanade	-AAM+Gabor+Adaboost+SVM system - voting processing to combine SVM outputs
[107]	SVM	79.9%	FEEDTUM	-4 expressions tested -classifier in cascade, combined with gender classifier
[160]	MLP, RBF, ANN, GLVQ	72%-MLP	Their own database	-leave-n-out strategy to train and test the classifiers -test: one person showing all facial expressions.
[115]	SVM	70.31%	FEEDTUM	-4 expression tested, 3 expression levels for each expression -train: 8 subjects, test: 2 unseen subjects, different image resolution
[162]	SVM (RBF)	71%	Image sequences from meetings	-presence of noise in the image sequences -incorporate 3-D information -train: 10-fold Cross-Validation

4.4. Facial Expression Recognition

While in expression classification, i.e., a binary class problem, we seek to choose between two classes of expression, in ***expression recognition,*** i.e., a multi-class problem, we want to be able to associate a human face with one particular facial expression.

Now, there are two main sets of classes used in the literature for automated facial expression recognition: the action units defined by Ekman and Friesen [40] and the prototypic facial expressions defined by Ekman [42]. Both concepts were previously explained in 2.3.1. In the same section we also discussed the differences between the two approaches.

Most researchers tend to directly interpret facial expressions in terms of the prototypic facial expressions. Only a few systems use rules in order to translate coded facial actions into expression classes. In our work, we are also going to interpret facial expressions in terms of the prototypic facial expressions (more details are given in Chapter 8).

As in the case of expression classification, different types of classifiers can be used. Basically, expression recognition can be seen as multiclass expression classification. Binary classifiers can be extended to accommodate multiclass classification. We are going to pay special attention to SVM, as it proved one of the most successful expression recognisers both in the literature [150] and in our practical experience.

4.4.1. Multi-Class SVM

By their nature, SVMs are intrinsically binary classifiers. However there exist strategies by which SVMs can be adapted to multiclass tasks. Two of the common approaches are One-Against-One (1A1), One-Against-All (1AA), Directed Acyclic Graph (DAG), and the cascade techniques.

The 1AA approach represents the earliest and most common SVM multiclass approach [163] and it involves the division of an N class dataset into N two-class cases. In the case of expression recognition, for the seven types of facial expressions, i.e., the six universal expressions plus the neutral one, we apply seven classifiers of the type expression/ non-expression.

The 1A1 approach [159] involves constructing a machine for each pair of classes resulting in N(N-1)/2 machines. In the case of recognising expressions, as N is seven, corresponding to the six universal expressions plus the neutral one, we will have 21 classifiers of the type expression1/ expression2. When applied to a test point, each classification gives one vote to the winning class and the point is labelled with the class having most votes. This approach can be further modified to give weighting to the voting process.

It is stated in the literature [164] that the disadvantage the 1AA approach has over 1A1 is that its performance can be compromised due to unbalanced training datasets. A practical example of an unbalanced training dataset is the case of consumer pictures when we have a reduced number of disgusted training faces, compared to a high amount of happy faces. The fact that people prefer to pose smiling rather than to have a disgusted face is non-debatable. However, the 1A1 approach is more computationally intensive since the results of more SVM pairs ought to be computed.

Another way to extend SVMs from binary classifiers to multi-class classifiers is by a ***Directed Acyclic Graph (DAG)*** [159]. $N(N-1)/2$ binary classifiers are created. Each binary classifier forms a node in the graph structure. Nodes are organised in the form of a triangle with the single root node at the top and increasing subsequently in an increment of one node in each level, until the last level that will have N nodes.

The DAG method eliminates one class out from the list at each level. At the root node, all classes are in the list. Each node discriminates between the first class and the last class in the list. Each level gives the result of one class out of the two classes; the class that is not in favour of that level is eliminated from the list. The procedure is terminated when only one class remains in the list.

An alternative option is to use classifiers for each class in a ***cascade structure*** [107]. Six classifiers, one for each basic expression, are arranged in a cascade structure. Each classifier is added iteratively, based on its binary classification performance. We are also going to utilise this approach in our experiments.

We conclude this section, by summarising some examples of results obtained in expression recognition, using a multi-class SVM approach.

Li et al. [150] propose a system that consists of four modules: a warp processing module using AAM, a Gabor filtering module, an AdaBoost training

module, and the SVM training and testing module. The SVM outputs are combined to make a seven alternative forced choice, by using voting to combine SVM outputs for multiclass decisions. Typical recognition rates of 90% were achieved.

Saatci et al. [107] used a cascade of SVM classifiers to perform expression recognition. Recognition rates improved when the expression classifiers cascade where preceded by a gender classifier.

In [162], Liebelt et al. use an SVM to recognise expressions on images of low quality or occluded. The authors obtain on average a recognition rate of 71.3%.

4.5. Databases used in the Expression Modelling and Recognition

In this section, we present databases used in the expression modelling community. We enumerate all the relevant databases that we are aware of, as it will be useful to have such a study in the literature.

Several databases have been developed for facial expression analysis. Despite various options, no single database is shared by all facial-expression research communities. Each database has its own particularities, e.g., different picture quality, lighting conditions, contrast, number of subjects, modalities to stimulate a subject's emotions, and so on. The lack of consistency between databases means that there is no single standard against which different algorithms can be directly compared. In a sense this is not surprising as it has taken many years for the research community to standardise on a number of large face recognition databases [165, 166]. As the field of expression recognition is more recent we expect that the research community is less mature and research standardisation less well developed.

Few databases contain *profile faces*. Probably the most important such database is the MMI database [167] (samples from this database can be seen in Figure 4.10).

Figure 4.10. Faces displaying facial expression of fear from profile and frontal views, as represented in MMI database [167].

Another issue is the *methodology of collecting suitable data* for the databases. Artificially posed expressions and facial expressions based on true underlying emotion present discrepancies, in both contracted muscles and in their dynamics [118].

Fasel and Luettin [39] acknowledged that posed facial expressions tend to be exaggerated and easier to recognise, as opposed to spontaneous expressions. This observation is supported by the fact that there are two distinct neural pathways that mediate facial expressions, i.e., one for volitional facial movement and one for involuntary facial movement [118]. Each one originates in a different area of the brain. One would expect that facial expressions mediated by these two pathways would exhibit differences both in the facial muscles which are involved and in the resulting dynamics of the formed expression. One example that supports this idea is the lack of constriction in the obicularis oculi[1] during artificial smiles [168].

Figure 4.11 displays examples of the universal facial expressions from two databases. In the first case, the Cohn-Kanade database [3] contains pictures of ***posed expressions***.

Figure 4.11. Examples of posed (first row) and spontaneous expressions (second row) from the Cohn-Kanade database (first row) and the MMI database (second row).

Before performing each display, an experimenter described and modelled the desired display. In the second case, there are presented examples of pictures from the FEEDTUM Facial Expressions and Emotions Database [169, 170]. The database belongs to the Technical University of Munich and it consists of elicited ***spontaneous expressions*** of eighteen subjects of the six universal expressions and the neutral one. To elicit the emotions by as natural a means as possible, the authors decided to play several carefully selected stimuli videos and to record the participants' reactions. For

1 The obicularis oculi is the muscle that closes the eyelid

this purpose a video monitor together with a mounted camera on top were employed, which enable a direct frontal view. Both devices were controlled by a dedicated software that induced the desired emotions by playing the recordings at specified times.

Table 4.2 contains a summary of the existing databases used by the expression recognition community. Their short description together with other characteristics is also enumerated.

TABLE 4.2. EXISTING DATABASES FOR EXPRESSION ANALYSIS TOGETHER WITH THEIR DESCRIPTION.

Database	Description	Other factors
CMU- Cohn-Kanade [108]	Video sequences of several subjects acting on multiple scenarios; each sequence includes a subject showing a specific basic facial expression	Expression intensity- from the neutral state to the apex of the emotion.
JAFFE - Japanese Female Facial Expression [171]	213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models.	Each image has been rated on 6 expressions adjectives by 60 Japanese subjects.
MMI [167]	2000 videos and 500 images of about 50 subjects displaying various facial expressions on command.	Metadata (displayed AU) is available. Frontal and pose views.
FABO -Bimodal Face and Body Gesture [172]	bimodal face and body expression database	Recorded simultaneously by face and body cameras.
Karolinska Directed Emotional Faces set [173]	490 colour pictures showing 70 subjects displaying the 7 basic expressions, 5 different angles	Posed database, excludes beards, moustaches, earrings, eyeglasses, and visible make-up.
eFEC database [174]	600 facial expression images 258 male and 342 female subjects, with 144 ‘neutral’, 124 ‘sleepy’, 177 ‘confuse’ and 155 ‘smile’ images; frontal view; inconstant lighting minor head rotation and orientation;	Glass/moustache; age within 20-35; multi-races/skin colour; with/without hair covered by cloth.
UT Dallas [175]	For each person there are 9 static “facial mug shots” and a series of video streams (moving facial mug shot, a facial speech clip, dynamic facial expression clips, 2 gait videos, and a conversation video)	Controlled lighting conditions; spontaneous database.
RU-FACS Spontaneous Expression [118]	100 subjects asked to either tell the truth or take opposite opinion and convince an interviewer they are telling the truth.	Rigorous FACS coding; spontaneous database.
Yale Faces database [176]	90 face images for 15 individuals; 6 different types of facial expressions: normal, happy, sad, sleepy, surprise and wink	Only small change in head pose and facial expression.

Purdue AR database [177]	4000 colour images corresponding to 126 people's faces (70 men and 56 women). The 4 expressions are neutral, smile, anger and scream.	Frontal view faces, various illumination conditions, occlusions.
POFA- Pictures of Facial Affect [178]	110 photographs of facial expression	Black and white pictures.
JACFEE/JACNeuF -Japanese and Caucasian Facial Expressions of Emotion [178]	56 photographs of different people, half male and half female, half Caucasian and half Japanese, illustrating the 6 universal expressions	Colour pictures, limitation: small number of pictures.
CVL Face Database [179]	114 persons (108 male, 6 female) in different conditions: profile left/right, 45 degrees left/right, frontal, frontal smile, and frontal smile with teeth	Colour pictures
FG-NET Aging database [180]	face images showing a number of subjects at different ages	-
FEEDTUM- Facial Expressions and Emotions [169, 170]	18 subjects within the MPEG-4 emotion-set plus added neutrality	Spontaneous database
Cohn and Kanade's DFAT-504 dataset [108]	100 university students ranging in age from 18 to 30 years. 65% were female, 15% were African-American, and 3% were Asian or Latino	Includes pose variations
Pictures of facial affect [181]	110 images consisting of six male and eight female subjects performing the 6 basic expressions	-
RaFD- Radboud Faces Database [182]	72 models showing expressions of happy, angry, sad, contemptuous, disgusted, neutral, fearful, surprised	Subjects of all ages, 3 different gaze directions and 5 different camera angles simultaneously.
MSFDE- Montreal Set of Facial Displays of Emotion [183]	144 facial stimuli expressed by 3 cultural groups; 5 different levels of intensity for each expression.	Limitation: small number of pictures available

Chapter 5 - Facial Feature Modelling with Standard AAM

The main goal of our work is to be able to automatically analyse facial expressions. This task requires detailed facial feature models that accurately describe facial expressions. In the previous chapters we chose and described AAM as a technique for whole-face description. AAM has been shown to work well when modelling whole faces [74].

In this chapter we will adapt the AAM technique in order to develop robust models for the eyes and lips regions of the face as these are the most significant facial sub-regions which determine facial expression.

We begin our approach by verifying the performance of the standard AAM formulation when modelling eyes and lips. We then adapt the AAM in order to model these challenging facial features. The eye model should depict not only the eye appearance, but also states like the degree of closure of the eye or the angle of eye gaze. Finally, a lips model should be able to handle challenges, such as the complex geometric variability of the lips and the varying textures of this region as the mouth opens and a person's teeth are exposed.

This chapter is divided into two main sections: in section 5.1, we apply and adapt AAM for the eye region. In section 5.1.1 we start by defining several terms related to eye actions and to our eye model. Then, in section 5.1.2, we summarise relevant work presented in the literature approaching eye detection and tracking, and blink and gaze detection. Section 5.1.3 provides a step-by-step description of the creation of the eye model, while section 5.1.4 presents some preliminary results and draws conclusions regarding our initial eye model.

In the second main section of this chapter, we use AAM to develop a lips model. Firstly, in section 5.2.1, we analyse and compare state-of-the-art in lips modelling. Then, we adapt the standard AAM for the lip region, as described in section 5.2.2. We present preliminary results and draw conclusions in 5.2.3.

We conclude the chapter in section 5.3, by analysing our initial models. We study their performance and describe their remaining challenges.

5.1. AAM Eye Model

5.1.1. Definitions of Main Actions Performed by Eyes

Several terms related to eye actions used in the current chapter are defined bellow, in order to ease reading and understanding of our work. Most of the definitions are cited from [184].

Eye movements have been described and analysed thoroughly by scientists from various domains. The eyes blink, wink, flutter, gaze, and perform saccades[2]. These actions bring relevant information about one's emotions or medical conditions [6, 185-188].

Blinking is defined as the rapid closing and opening of the eyelid. This involuntary action has several health benefits, as it spreads tears across the eyes and it removes irritants from the surface of the cornea and conjunctiva. A person approximately blinks once every five seconds [189]. Infants do not blink at the same rate of adults; in fact, infants only blink at an average rate of one or two times per minute [190]. On average, a blink takes approximately 300 to 400 milliseconds [189]. Blink speed can be affected by elements such as fatigue, eye injury, medication, and disease [191]. In consequence, it can be a criterion for diagnosing drowsiness states [6, 185], emotions [186], or medical conditions [187, 188], e.g., blink rate decreases in Parkinson's disease [187] and increases in schizophrenia [188].

A few instances of multiple blinks in a short period of time are called ***eyelid flutter***, defined as two or more rapidly repeating blinks.

When a person only blinks with one eye, it is called ***winking***. Winking is usually voluntary and it can be used as a form of body language, communicating agreement and affection.

Gaze is defined by the direction the eyes are pointing in.

On a more general note, eye movements can be divided in ***fixations***, i.e. when eye gaze stops at one position, and ***saccades***, i.e., when eye gaze moves suddenly to another position. Fixations are eye movements which stabilise the retina over a stationary object of interest. Saccades are involuntary rapid changes in the point of fixation, during scanning of objects, reading, etc. Saccades are the fastest movements

2 Eye saccades are quick, simultaneous movements of both eyes in the same direction.

that human body can perform. During the day the eyes can rotate over 500 degrees per second, and on average 1000 of these saccades are performed [192].

Many automatic systems have been developed to analyse, model, and track these eye movements. As an initial step the eye region must be detected in a facial image. ***Eye detection*** is challenging due to the weak contrast between the eye and the surrounding skin [28].

In some applications only ***pupil detection*** is required [193]. The pupil is the black circle in the centre of the iris (see Figure 5.1); it carries important information, e.g., the size of the pupil determines the amount of light that enters the eye.

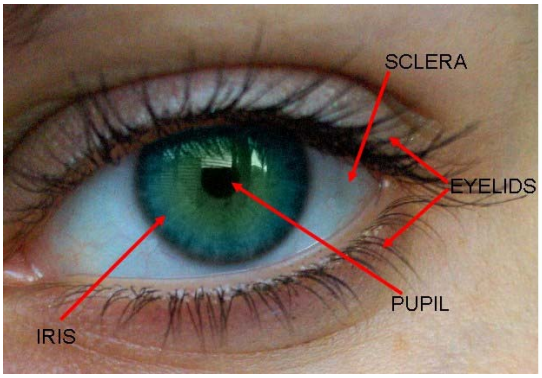


Figure 5.1. Description of human eye.

A related application is eye tracking. ***Eye tracking*** is the process of measuring the gaze, eye position, and their motion. An ideal eye tracker system [28] should be robust, non intrusive, i.e., avoid any physical contact with the user, and inexpensive.

Eye detection, along with eye and gaze tracking, received a great deal of attention for applications such as facial expression analysis [186], computer animation [194, 195], driver awareness systems [6, 185], film and advertising industry [196], or communication interfaces using blink and gaze for people with severe disabilities [197, 198].

In the following section, state-of-the-art eye-related techniques are described and compared. We discuss research work in eye detection and tracking, blink detection, gaze detection and tracking.

5.1.2. State-of-the-art in Eye Modelling

Eye-related techniques can be classified into two categories: ***intrusive***, which require direct contact with the eyes, and ***non-intrusive***, which avoid any physical contact with

the user. The latter is more of practical interest for consumer applications, as the user is not constrained by uncomfortable equipment.

A good survey of non-intrusive methods is offered in [28], that broadly classifies non-intrusive methods into three principle categories: feature-based [19], appearance-based [28, 199-202], and model-based [197, 198, 202-205]. Each category is going to be described below. We only describe methods based on active illumination, ignoring the methods based on infra-red (IR) illumination [28, 193].

5.1.2.1. Eye Detection and Tracking

Feature-based methods are based on tracking individual features of the eyes, by exploring their characteristics, such as edge and intensity of iris, eye colour, or eye corners.

A popular method is the extraction and tracking of the eyelids [203], robustly under various illuminations and face orientations. Firstly, the authors of [203] obtain a rough location of the eyes by morphological filtering. This type of filtering permits to extract contrasted components of a given size. The dark contrasted components, i.e., pupils and eyelids, are combined with the white contrasted components, i.e., the white of the eyes, defining the eye region. Then a dynamic programming algorithm, i.e., the minimal path algorithm, is applied to extract the precise shape of the eyelids.

In [206], instead of detecting the eyes, the authors detect the mid-point between the two eyes, based on pixel-value characteristics. The pixel values of this point has two circles of bright parts, i.e., forehead and nose bridge, and of dark parts, i.e., eyes and brows. A circle-frequency filter that has a local maximum at these characteristic points is applied. Eyes are subsequently detected as two dark parts, symmetrically located on each side of this mid-point.

In ***appearance-based*** methods, the model is learned from a large set of original images, so it does not need to build further models for objects and no additional features need to be extracted.

In [35], eigenfaces (section 2.2.2.1) are extended for the eyes. The authors of [35] extract an appropriate eye template for training and they build a principal component projective space called ***eigeneyes***. Eye detection is accomplished by comparing a query image with an eye image in the eigeneye space.

In [202, 207] the motion information between two consecutive frames is considered to detect the eyes. This motion is likely to be caused by an eye action, hence providing information of the eye location.

In ***model-based*** methods, firstly a generic eye model is designed and then a template matching is used to search the eye images. Deformable templates methods [202, 203, 205, 208] are commonly used in this case. These approaches usually require a good initialisation for the matching algorithm, i.e., accurate eye detection.

In [209], Lam et al. firstly detect the eyes by means of geometric measures. Eye corners are detected and used to set the initial parameters. Based on this initialisation, the authors locate the eye template in relation to the eye images, greatly reducing the processing time.

In [210], the eye is modelled as a multilayered parameterised model that consists of multiple components corresponding to the anatomy of an eye: iris, upper and lower eyelids, white region around the iris, bulge, and furrow. A model for each component is rendered in a separate rectangular layer. When overlaid, these layers represent the eye region. Tracking eye motion is done by matching the eye region.

AAM [208] represents a popular and robust ***hybrid*** technique combining model-based and appearance-based techniques for eye tracking. Its main advantage is that the shape and texture are exploited directly as they appear in the original image. The method is also invariant to rotation, scale, or pose. In [208] there are described limitations regarding the placement of the recording camera. It has to be below the line of sight of the user because it needs to be placed fairly close to the subject, so it will not obscure the user and will have a better view of the eyes.

An extension of AAM is presented in [211], where the eye region is divided into three sub-regions: upper eyelid, lower eyelid, and palpebral fissure, i.e., separation between the upper and lower eyelids. An AAM is trained for each. The method firstly acquires the intrinsic structure of the eye in the initial frame where the eye is known to be neutral, and then tracks the motion across the entire image sequence.

5.1.2.2. Eye Blink Detection

Current eye models function well for open eyes, but they are not robust against physiological changes in eye appearance, such as blinking. A determination of the eye state, i.e. open or closed, is more complex than simply determining the eye location.

This is due in part to the small size of the eyes relative to the face region and to the weak contrast between the eye region and the surrounding skin [212].

Numerous methods to detect the blink have been proposed. In this discussion we list some of the more recent work, from a considerable body of literature. Intrusive techniques have also been proposed, e.g., analysing features in electro-encephalograms (EEG) [213] and electro-oculographic potential (EOG) [214], but we only focus on non-intrusive techniques.

From the ***feature-based*** category, we mention blink detection based on iris tracking [202]. It is based on the hypothesis that the disappearance of the iris implies a blink. The eye is tracked and constantly monitored to establish to what extent it is open or closed at each frame. Blink duration is defined as the count of consecutive frames of closure.

Using images from a high-resolution video-camera, Moryiama et al. [199] describe an ***appearance-based*** approach. The authors divide the eye-region in two: upper and lower half, and calculate the average illumination intensity of these two halves. They noticed that the mean intensities for the two halves plotted over time cross when eyes blink. Therefore by counting the number of crossings they detect the timing, number, and duration of blinks.

Blinking can also be detected from frame differencing by detecting the motion between two consecutive frames [201]. In this approach head movements have to be taken into consideration as well as eye positions, leading though to a quite complex analysis.

In [200] the second order change, i.e., eye blink, is discriminated from a first order change, i.e., head movement. After the pixels corresponding to the head motion have been filtered out by the second order change detection using three consecutives frames, the remaining motion pixels are used to detect a blink.

Template matching techniques, belonging to the ***model-based*** category, have also been proposed [198, 202]. An example of such approach is presented in [202], where Grauman et al. calculate correlation scores between the actual eye detected by "blink-like" motion and a closed-eye template.

5.1.2.3. Eye Gaze Detection and Gaze Tracking

Gaze tracking can be performed intrusively, e.g., contact lenses [215], electro-oculography (EOG) [215], and non-intrusively [6, 193, 216, 217]. As in the case of

eye detection, the latter category is of practical interest for consumer applications, as the user is not constrained by uncomfortable equipment. Also, many methods are based on IR illumination [193, 216]. In this section, we only exemplify methods based active illumination, as they correspond to our interest.

In [217], gaze tracking is performed using a contour-based tracker. This technique is based on the fact that the iris is circular and characterised by a large contrast to the sclera. A dynamic model of state propagation is combined with an observation model determining the iris contour from the image data.

Magee et al. [218] determine the gaze direction by exploiting symmetry between the left and right eye. Firstly, the authors use frame difference to determine changes between frames. When a change has been detected, the left eye is mirrored and subtracted from the right eye. If the user is looking at the camera, this difference will be very small; if the eyes are looking towards left, the mirrored left eye will be looking right, etc. The authors emphasise that their algorithm is able to concurrently run with the real-time vision interface systems.

In [6], AAM is used to track the whole head, enabling the location of eye corners, the eye-region, and the head pose to be determined. Then, a geometric model estimates the gaze, as described in Figure 5.2. The authors conclude that the robustness and accuracy of the algorithm are actually derived from the AAM tracking of the whole head, rather than from the gaze detector. Once the eye corners have been located, finding the irises and computing the gaze are straightforward.

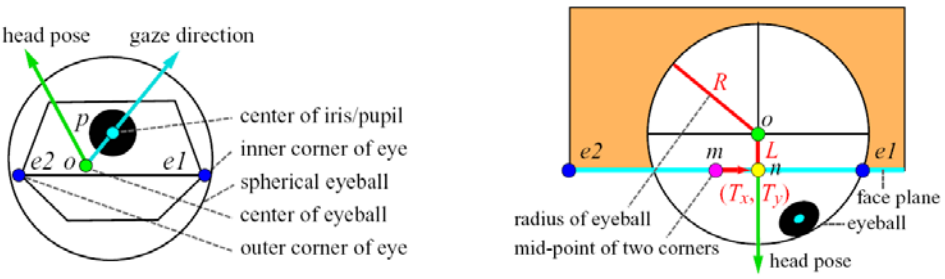


Figure 5.2. The gaze estimation geometric model, as proposed in [6].

5.1.2.4. Discussion

The aim of our work is to develop an eye model that will serve several consumer applications, e.g., blink detector, eye tracker, facial expression analyser. In consequence, in our approach we aim for non-intrusiveness and naturalness.

We chose AAM to develop a complex eye model, robust to blink, wink, and gaze. Our approach belongs to the model-based category, but ideas from the

appearance-based methods are also incorporated. The main advantage of employing AAM is that the shape and texture of the eye are modelled directly as they appear in the original image. Compared to other approaches, our approach is also invariant to rotation, scale, or pose and it is capable of successfully synthesising eyes in unseen pictures.

Although AAM techniques were used by Hansen et al. [208] and by Moriyama et al. [211] to model the eye region, these techniques have not yet been applied to model eye blinking. We propose a solution that models eye blink, and can be readily adapted to determine eye-gaze and wink. This solution is described in the following sections.

5.1.3. The AAM Eye Model – Initial Implementation

In this section, we describe an initial implementation of our eye model, using the standard AAM formulation. The model we have developed offers a detailed analysis of the eye region in terms of degree of eyelid opening, position of the iris, and shape and texture of the eye. The model is designed to be robust to small pose variation.

Blink and gaze models are subsequently developed.

The reader will note that the techniques used in this first statistical appearance model of the eye region are mostly drawn from our earlier work with holistic face models. There are, however, some novel refinements – in particular the unique use of overlapping landmark points to model eye-open and eye-closed states and a range of intermediate degrees of the eye being closed.

Now in section 5.1.3.1, we begin by describing in detail the way in which AAM is employed to model eye appearance. We detail every step in the model building, rationalising our choices in shape and texture modelling.

Both open and closed eye appearances are modelled, in order to accommodate the eye-blink, as described in section 5.1.3.2. In the same section we explain how we design the model to make it robust to small pose variation.

Finally the model is extended to accommodate variations in eye-gaze. This extension is described in detail in section 5.1.3.3.

5.1.3.1. Statistically Modelling the Eye Appearance

The appearance of the eye-region is represented by a statistical model of appearance, following the procedure described in section 3.2.

As explained in section 3.2.1, an important decision when building a statistical model of appearance is the choice of the region to model and of the landmarks that demarcate the shape: their number and their positioning. Depending on the required level of detail and on the facial image resolution used, the number of landmarks can vary. Their choice depends on the application, the answer being a compromise between the number of parameters, the fitting accuracy, and speed.

Various numbers of landmarks and ways of positioning them were initially tested; some initial annotation examples are given in Figure 5.3. From the preliminary experiments it was determined that the eyebrows are a key feature for accurate eye detection, as they bring significant information related to eye position. The iris is also an essential feature; it allows us to incorporate eye-gaze in addition to eye-blink.

Our model incorporates both eyes. It is known that left and right eye typically blink in a closely synchronised fashion. Thus it should be sufficient to capture the trace for one eye in order to measure eye-blink for a single subject. However with a detailed two-eye model, we can include additional information and distinguish other eye states, such as a wink.

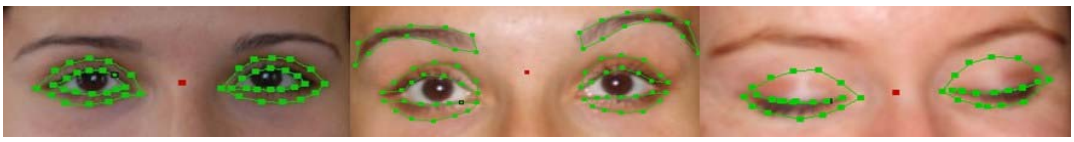


Figure 5.3. Different types of annotations: from left to right, 48, 72, and 38 landmarks. The first annotation lacks a precise eye location, caused by the absence of the eyebrows; the second annotation is not robust to eye gazing, caused by the absence of the irises; the third annotation is only valid for closed eyes.

Subsequent to these experiments, we chose to annotate eyes as shown in Figure 5.4. The model is composed of the following elements: eyelids, eyebrows, and irises. The model annotation is described by a number of 80 points, disposed as to capture the most important eye features. The model is unique compared to the literature in that it is valid for both open and closed eyes.

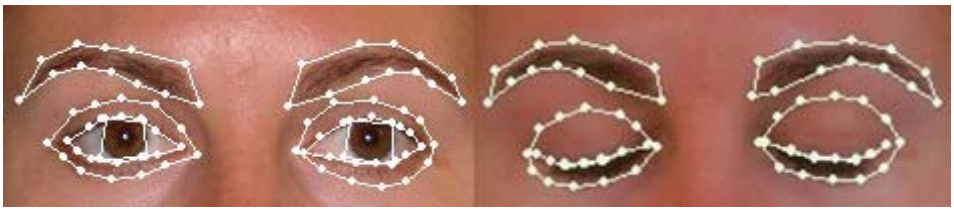


Figure 5.4. Image annotations for open and closed eyes.

5.1.3.2. Statistically Modelling the Blink

One of the main requirements of this eye-model is to robustly model the eye blink. A blink sequence can be described by the following sequence of actions: fully-open-eye, closed-eye, and fully-open-eye. In order to include blinks, the model has to be able to describe this movement.

Including blink appearance

In order to model eye-blink, we have extended the standard model to incorporate within the same model two different templates, for open and closed eyes. A consistency between the two templates needs to be found.

As explained in section 3.2.1, the number of landmarks used for annotating the shape has to be kept fixed over the training data set. The landmarks should predominantly target fiducial points, which permit a good description of facial geometry. Thus, one should choose a number of landmarks that best characterises eyes over the entire training set, whether the eyes are open or closed. One should pay attention to the fact that some of the facial details might not be visible in all training images, e.g., irises are occluded when eyes are closed.

We annotate the variability between open and closed eyes through using ***the same number of points*** for the two templates, ***mapped in the same order***, where the upper and lower inner eyelids overlap. This overlapping is represented in Figure 5.5, by the red arrows: points 9, 10, 11, 12, and 1 overlapping with points 7, 6, 5, 4, and 3 respectively. And in the second eye points 33, 34, 35, 36, and 25 overlap with points 31, 30, 29, 28, 27. The procedure emulates the physiological blinking action, when the upper eyelid closes over the lower eyelid.

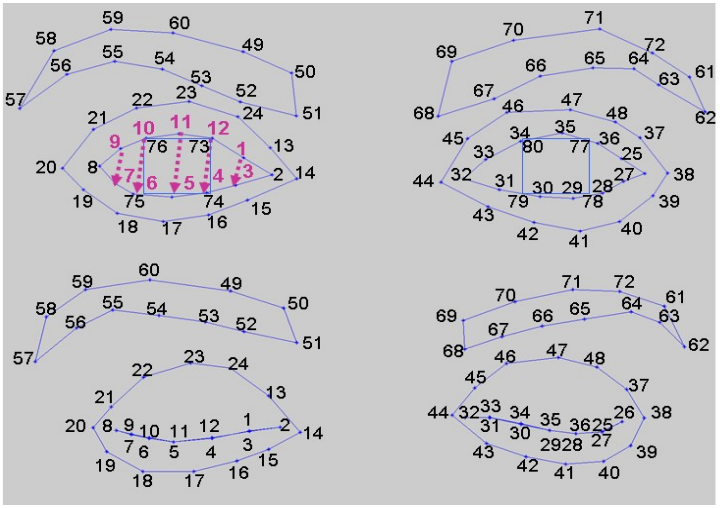


Figure 5.5. The two annotations for the open/closed eyes; the latter annotation is obtained by overlapping the points from the upper inner eyelid.

Separating the blink parameters

Two different training sets were used in the training procedure, one for open eyes and one for closed eyes. The model is able to represent all the states in between, e.g., semi-closed eyes, even though they are not included in the training sets. This can be explained by the fact that AAM is a deformable model, capable of adapting to the variation in both shape and texture between open and closed eyes.

In these experiments we use 30 pictures for each training set. Examples of these training images are given in Figure 5.6. As we want to develop a ***general*** model (section 3.5), we include a high degree of variability in these training pictures, in terms of subject individuality, illumination, etc. The resulting AAM model has 25 shape parameters and 33 texture parameters.

Figure 5.6. Examples of training pictures for open and closed eyes.

Since we want to be able to separately process shape and texture parameters, we use an independent AAM (section 3.2.3). We observe the values of shape and texture parameters and we note that there is a differentiation between closed-eye and

open-eye parameters, as seen in Figure 5.7. The figure displays an average over the parameters with the most important variation of texture, in the first graph, and shape, in the second graph.

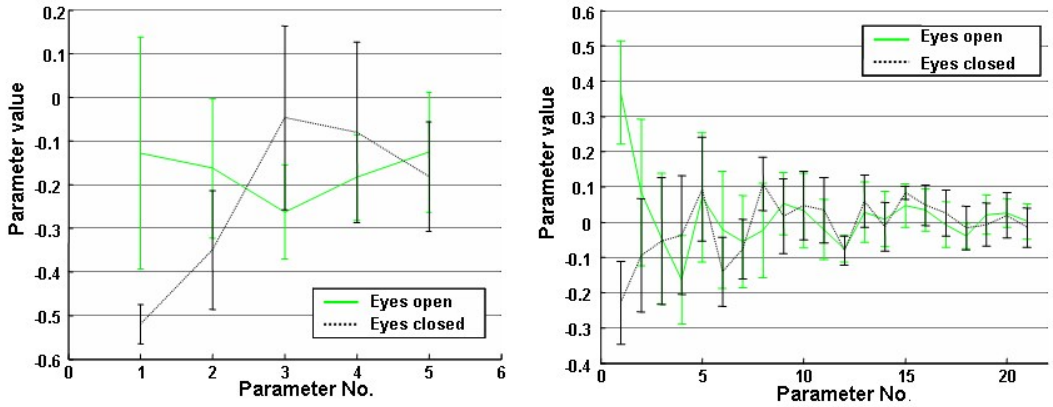


Figure 5.7. Average values and the unit of standard deviation for texture (first graph) and shape parameters (second graph) tested on images with eyes open (green curve) and closed (black curve), respectively.

The differentiation is more visible for the shape parameters. Figure 5.7 indicates also that the differentiation is more important for the parameters encoding the higher variation, i.e., the first two parameters in the case of shape and the first parameter in the case of texture parameters. These parameters correspond to the blink action.

From our earlier experiments and work with the AAM face model this is as we expected. The variation due to blink is decorrelated from other sources of shape variability and is also the most significant of the variations encountered in the training set.

In section 4.2.2.1 we explained that although PCA jointly models all sources of variability [130], it gives initial estimates of the sources of variability [119]. We visually exemplified the parameter separation in Figure 5.7. The differentiation depends on the complexity of the training set. For example, we observed that if the training set contains only frontal faces with open eyes, the biggest variation is attributed to the difference between individuals. By careful selection of the training image set the model differentiation can be selectively chosen to emphasise gaze parameters, pose, or shape-of-the-eye parameters. We will experiment with some of these later in this section.

Including pose variation

A problem which generally appears in face modelling is that of robustness to pose. As we proceeded for full face modelling in [219], the eye model can be designed in order to include head-pose variation within a limited range of angles, around the horizontal and vertical axes. Decoupling the pose variation from the global shape model can be realised by using a specialised training set, in which the individuals are presented in several poses, normally covering a range within 30-40 degrees for head tilting.

To address this issue, we included additional head pose variations in the initial training set, Figure 5.6. 10 pictures containing pose variation were added to the initial training set. We obtain 27 shape and 39 texture parameters. Thus, the model is enriched with two new shape parameters, accounting primarily for pose variation.

Figure 5.8 shows the effect of varying with +/- one unit of standard deviation for some representative shape parameters of the model.

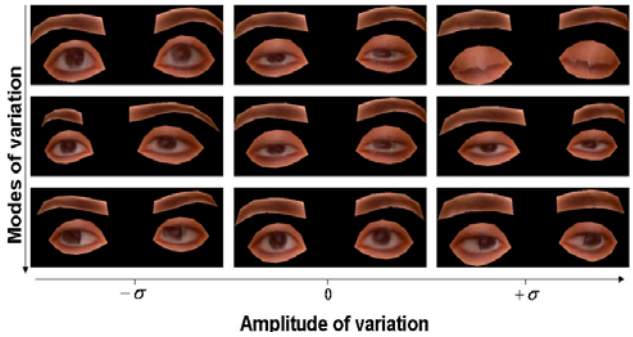


Figure 5.8. Modes of variation for the shape parameters by +/- one standard deviation from the mean. The variation from the mean of the blink parameters is represented in the first row, while the variation of pose and gazing are represented in the last rows.

The equation (3.1) describing an AAM general shape model can be extended to:

$$s = \bar{s} + \varphi_{blink} b_{blink} + \varphi_{pose} b_{pose} + \varphi_{gaze} b_{gaze} + \varphi_{eye\_shape} b_{eye\_shape} \quad (5.1)$$

in the case of our eye model. The equation explicitly separates shape parameters that account for the different sources of variation, such as blink, gaze, and different eye shapes. The types and number of each type of variation depend on the number of training pictures describing this variation.

5.1.3.3. Statistically Modelling the Gaze

The gaze model is based on information provided by the statistical model describing the eye region described in section 5.1.3.1. The original model, while developed for

blink, is actually detailed enough to deal with other eye actions, such as eye-gaze. The model represents information regarding size and colour of iris, width and boldness of the eyelid, iris 2-D position, etc.

Based on the same principle as in 5.1.3.2, by including in the training set pictures describing gaze variation, the model will learn this variation. Then it will be able to model the gaze variation in unseen pictures. For our model, pictures containing gaze were initially introduced into the training set.

As we already experimented and explained in section 5.1.3.2, equation (5.1) separates shape parameters modelling the gaze. In the case of our training set, two shape parameters account for gaze variation.

The presence of the iris in the model and the knowledge about eye-corners give us enough information to describe the gaze. The position of the iris centre moves when the gaze direction moves. The variation in the eye regions for various extremes of the gaze shape parameters is shown in Figure 5.9.

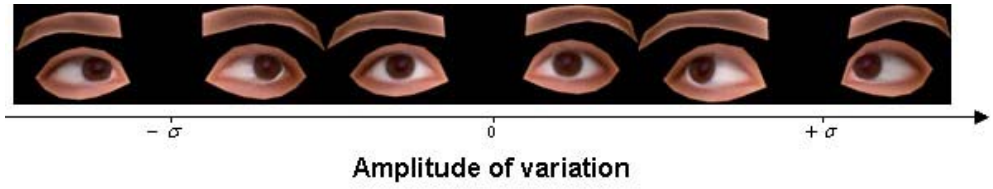


Figure 5.9. Modes of variation for the eye model by +/- one standard deviation from the mean, described by the gaze shape parameters.

5.1.4. Results and Conclusions for our Initial Eye Model

Preliminary experiments were performed, in order to analyse the initial AAM eye model. The requirement at this stage is that both open-eye and closed-eye shapes are accurately fitted.

We tested the model described in section 5.1.3 on 60 consumer pictures of various subjects, with open or closed eyes. Some of the pictures also included pose and illumination variation. None of the test pictures were included in the training set. At this stage, we only qualitatively analysed the results. After visual inspection, the fitting results were considered as failing in over 60% of pictures indicating a lack of generalisation of the model.

Some results can be judged from Figure 5.10. In the first row of Figure 5.10, we exemplify poor shape fitting in case of facial expressions, e.g., surprise, anger. The second row contains poor fitting in case of pose variation. Although we

integrated pose in the training variation, the model is still sensitive to even slight variation from the training poses.

Illumination variation generally affects facial modelling, and in particular, affects our eye model. Examples of failures caused by unseen illumination are displayed in the third row of Figure 5.10. Other causes of poor fitting are occlusions, such as glasses or fringe, red-eyes, or wink-eyes. In the last row, we exemplify situations when open eyes are matched with closed ones and closed eyes are matched with open ones. These failures are caused by various factors, such as gaze or head pose.

Figure 5.10. Examples of shape fitting for our AAM eye model.

Our initial AAM model includes open and closed eyes, small pose and gaze variation. When tested, we acknowledged the fact that in its current state the model fails many challenges, as the ones exemplified in Figure 5.10. The conclusions motivate us to improve our eye model to make it more robust to the listed challenges.

An improved version of our AAM eye model will be presented in Chapter 6.

5.2. AAM Lip Model

In the previous section, we adapted AAM for the eye region. Another facial feature of interest is the lip. This facial feature, well known as being the most deformable facial feature, represents a major challenge in face modelling. In the current section, we test the feasibility of lip modelling using the standard formulation of AAM.

This section is organised as follows. Firstly, in section 5.2.1, we summarise relevant work in literature aiming to model the lip region. We indicate the main categories of approaches, while discussing the remaining challenges. Then, in section

5.2.2, we adapt the AAM approach for the lip region. Results and conclusions for the proposed model are presented in 5.2.3.

5.2.1. State-of-the-art in Lip Modelling

In literature, several methods have been proposed and they can be mainly categorised, as explained in [220, 221], as feature-based [29, 221-224], model-based [225-228], or combinations of the two, i.e., hybrid approaches [29, 229].

The first category, the feature-based approaches, is based on modelling individual lip features by exploring their characteristics, such as lip colour. Lip extraction is primarily based on colour-based segmentation techniques. Usually the processing is done in the Hue, Saturation, and Value (HSV) space [29, 224]. The difficulty of using the RGB space is that it is hardware oriented and is suitable for acquisition or display devices, but not particularly applicable in describing the perception of colour [230]. Important advantages of HSV are the separability of chromatic values and the possibility of using only hue for segmentation. Usually, hue is invariant to shadows, shading, and highlights [224].

In [223] the notion of colour constancy is used. Colour constancy refers to the ability to identify a surface as having the same colour under considerably different viewing conditions. The authors use a Bayesian decision classifier using chrominance as feature for segmentation, to make it adaptive to different viewing conditions. The segmentation is further improved by including features which describe chromatic second order statistics of neighbouring pixels.

In [221, 222], segmentation of the colour lip image is achieved by spatial fuzzy clustering.

Model-based techniques, or ***holistic*** approaches, are based on template matching to search the lip images. The approaches include Deformable Templates [225, 226], Active Shape Models [227, 228], Active Appearance Model [231, 232], and Snakes [228].

In [232], Matthews et al. compared the use of AAM, ASM, and multi-scale analysis (MSA) in lip reading, concluding that AAM achieved better recognition rates than ASM. MSA and AAM yield similar fitting rates, but they fail in different ways.

In parallel with our research, Kang et al. [231] uses AAM to model lips. Due to poor AAM fitting rates, an improved training method for AAM is proposed to

reduce the manual error of landmarks in the training set, based on iteratively adapting the landmarks by repeated fitting.

Liew et al. [226] state that the use of a parametric lip model as a template has the advantage of being more robust since the allowable shape is pre-constrained. In addition, the lip shape is described using a small number of parameters and the parameters have clear physical interpretation.

Gacon et al. [227] propose a method based on a statistical model of shape. The cost function used to fit the model is based on the response of Gaussian local descriptors. Knowing the shape, the authors predict the response of the descriptors with a non-linear neural network. Their approach is particularly adapted for the inner lip contour and mouth bottom which present a high variability and non-linearities.

Discussion

The feature-based techniques have the advantage of providing simple methods which work well under environmental constraints, but that lack the required robustness for real conditions [29].

Model-based techniques compensate for the drawbacks of the feature-based methods, but they are computationally expensive and dependent on the constraints learned from examples [29]. It is common to see non-optimal local alignments, when attempting to model an object that is quite different from the training set. It is especially noticeable when dealing with features that can fade into the background, such as the lips in our case.

In [29], Pantic et al. combine feature-based and model-based approaches in search for a robust mouth detector. The authors develop a ***hybrid*** knowledge-based technique for mouth feature extraction from facial images. At first, they use a feature-based approach. A coarse mouth detection is performed by applying a hue filter to filter the lips region, given that the natural colour of the lips is red. Then, two model-based algorithms, namely the curve fitting of the mouth and the mouth template matching, localise the mouth contour.

From the described work, we conclude that lips are considered a weak feature because of their complex expression and varying pattern, but also because of their weak colour contrast and significant overlap in colour features with the face region. Factors such as illumination and skin reflectance, or occlusions, such as teeth and tongue, complicate the task.

In our work, we firstly verify the ability of AAM to search for the lip region. In the current section, we describe the way in which we adapted AAM for the lip region. Then we test the model and discuss remaining challenges.

5.2.2. Description of our Initial Lip Model

A statistical model of appearance is built for the mouth region. The development of the model follows the algorithm steps described in section 3.3. We follow the same approach as for the eye model (section 5.1.3) in: choosing the training set, annotating shapes, and observing the separation of the shape parameters following different sources of variability.

Lip appearance is considered as the most variable facial feature [29]. It is conditioned by many factors, such as skin colour, make-up, illumination, head pose, mouth expression, or the presence of teeth and the tongue. As in the case of the eye model, our end goal is to be able to generalise our model in terms of individuals, pose, mouth open (teeth) and closed, and in terms of specific actions performed by the lips, etc. We achieve this by including in our training set pictures representative of the major factors of variability mentioned above. Some samples of our 30 pictures used in the training set are shown in Figure 5.11.

Figure 5.11. Examples of the training images for the lip model. The pictures include different degrees of teeth exposure (first row), various individuals with different colour/shape of the lips (second row), or different head pose and lip expression finalised in different degrees of lip curvature (third row).

The next step is shape annotation. If the annotation is optimal, the model will provide a detailed analysis of the lip colour, their degree of opening, their curvature etc. We empirically determined, by a trial-and-error approach, that a suitable annotation would be as the one illustrated in Figure 5.12. The figure shows our annotation that uses 24 landmark points.

The only requirement is to coherently annotate all images, using the same number of points mapped in the same order. Following the principle developed for the eye model (section 5.1.3.2), if the mouth is closed, the corresponding points from the inner lip contour are merged, i.e., points 20-23, and 12 overlap with points 14-18.

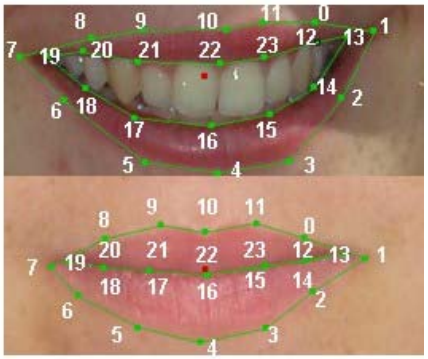


Figure 5.12. Mouth annotation for different expressions.

By training our lip model following the procedure developed in section 3.2, we obtain 12 shape parameters and 17 texture parameters. These parameters correspond with an allowed variance of 95% for both shape and texture modes. The shape parameters can be differentiated into parameters describing the opening of the mouth, smile, head pose, etc. The separation is a consequence of applying PCA on an educated choice of training pictures, i.e., pictures that include a balance of the relevant sources of variation. It is the same effect that we observed in the case of the eye model (Figure 5.8).

Figure 5.13 shows the effect of varying with +/- one unit of standard deviation some representative parameters of the model: texture (first row) and shape (last three rows). The shape parameters relate to (i) degree of mouth open/closed; (ii) up-down orientation of the mouth; (iii) left-right orientation of the mouth caused by head pose.

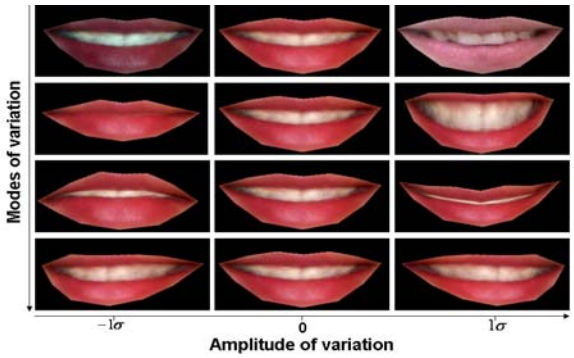


Figure 5.13. Modes of variation for the mouth model by +/- one standard deviation from the mean. The variation from the mean of texture parameters is represented in the first row, while the variation of different expressions and poses are exposed in the last rows.

5.2.3. Results and Conclusions of our Initial Lip Model

As in the case of the eye model, we test the lip model by verifying its matching to a group of unseen images. We used 50 pictures, including different individuals, facial expressions, e.g., happy, angry, and head poses.

The model initialisation is inferred from the face detection step. It uses the statistical relation between the rectangle describing the surroundings of the face obtained after face detection and the lip position in the rectangle learned from the training dataset.

In the case of our AAM lip model, after visually inspecting the results, we note that it fails to accurately fit unseen images in a proportion of 80%. The model faces the same challenges as the eye model, but with even poorer fitting rates. Figure 5.14 displays some examples of poor fitting.

Figure 5.14. Examples of poor fitting of the AAM lip model.

The main difference is that while an initial eye location is accurately determined based on face detection stage, the lip model frequently fails this first initialisation step. This is mainly due to the weak colour contrast and significant overlap in colour features with the face region. As explained in section 3.4, this initialisation error will adversely affect the final fitting results.

In conclusion, our lips model requires a more reliable initialisation step. In the next chapter we propose a solution to this challenge.

Chapter 6 - Extensions of the AAM Facial Feature Models

In this chapter, we propose extensions of our facial feature models, in order to solve the challenges of the standard AAM formulation and to make them robust to unseen facial variation. This approach, as we will shortly see, helps overcome the poor fitting of eye and mouth models to unseen images as shown in Chapter 5. An important part of the research presented in this chapter was published in [233-235].

This chapter is divided into two main sections: section 6.1 is concerned with the presentation of an advanced AAM eye model and section 6.2 develops similar improvements for the AAM lips model.

In section 6.1, we propose a component-based AAM adapted for our eye region model described in Chapter 5. In section 6.1.1 we initially propose two versions of this approach. Subsequent to their comparison, in section 6.1.2, we chose the approach that best suits the requirements of our eye model. The proposed model is then verified by means of some test applications, such as an eye tracker and a blink detector, described in section 6.1.3.

The second part of Chapter 6 is dedicated to an improved AAM lip model. As seen in section 5.2.3, the standard AAM lip model requires a strong initialisation to reduce the high rate of fitting errors (> 80%). In section 6.2.1, we propose a hue filter in order to obtain an improved initial location of the lips. In the following section, 6.2.2, we describe the overall lip modelling approach. As in the case of the eye model, we verify the lip model performance in section 6.2.3 by means of some test applications, such as a lip tracker and a smile detector.

Note that this chapter is concerned with the design and implementation of enhanced models; thus while we include some qualitative observations in this chapter the detailed testing and comparison of these models will be documented in Chapter 7.

6.1. Extension of the AAM Eye Model

The initial eye model developed in Chapter 5 is based on the standard formulation of AAM [74]. The model offers a detailed analysis of the eye region in terms of degree of eyelid opening, position of the iris, and shape and texture of the eye. The model is

designed to be robust to small pose variation, as described in section 5.1.3.2. Blink and gaze actions were additionally modelled in section 5.1.3.2 and, respectively, 5.1.3.3.

However, when we tested the model on general conditions not included in the training set, it failed in 60% of the cases, as we experienced in section 5.1.4. Its challenges include unseen head pose, occlusions of one or more components of the eye model, or difference in expression between the two eyes, e.g., winking.

Its challenges are explained by the limitations of a standard AAM. Although AAMs are powerful tools for image interpretation, they present several limitations when used as global appearance models. This formulation, despite all its advantages described in section 3.2, has the limitation of constraining the model to the variations learned during the training phase. As we created a global model for the two eyes together, we subsequently realised that this constraint does not allow the two eyes to deform independently.

Figure 6.1 visually explains this type of drawback. The picture presents two types of challenges: in-plane-head-rotation and independent actions of the eyes, i.e., winking. In the training set we included pictures with the two eyes open or closed. However, as we modelled both eyes together, the model cannot adapt to situations such as one eye open, while the other is closed.

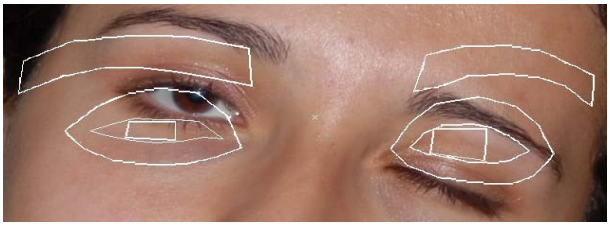


Figure 6.1. Example of poor fitting for winking in-plane-rotated eyes.

The ***component-based AAM***, as described in [94, 95], offers a solution to this particular drawback of AAM, i.e., constraining the model components to global variations. The authors of [94, 95] propose a face model that combines a global model with a series of sub-models. This approach benefits from both the generality of a global AAM and the local optimisations provided by its sub-models. We now adapt this approach to independently model the eye regions within a face.

In the next section, we propose two versions of the component-based AAM adapted for the eye region. We test both versions, in section 6.1.2, and decide for one

version that best suits our requirements. This particular version of the component-based AAM will be implemented for the eye region and used in different applications.

6.1.1. Component-based AAM Formulations

In a first stage, we adapted the component-based AAM for the eye-region using two different approaches. For the first approach the open-eye and closed-eye states are modelled separately; for the second approach we model each eye separately, retaining the mixed (overlapping points) open/closed information for each eye region.

Using these two distinct approaches we also hope to learn more about how AAM models behave under different training constraints.

Component-based eye AAM, separating open and closed eye states

In our first approach, we separately modelled the two-eye appearance for open or closed eyes. We started from the hypothesis that mixing open and closed eye shapes or textures can introduce errors into the model.

The procedure is as following. We firstly create a global model using both eyes, open and then with both eyes closed. The model is refined with two sub-models modelling only some particular features of both eyes, meticulously annotated as shown in Figure 6.2. A first sub-model represents components of open eyes, i.e., inner eyelid and iris. A second sub-model represents components of closed eyes, i.e., inner eyelid and outer eyelid.

In the case of closed eyes, the inner eyelid is not composed by overlapped points, but represented as a straight line. The eyebrows are included only in the global model for better eye location. They are considered superfluous for local modelling stages, as the eyebrows are mostly necessary for accurate eye location. In these stages, the eye location is believed accurate, coming from the initial global stage.

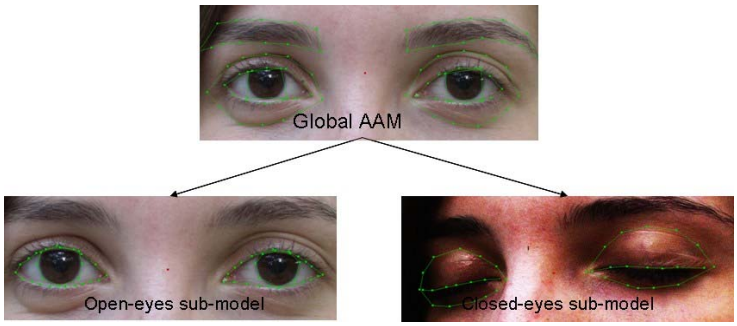


Figure 6.2. Annotation for the global AAM and for the two sub-models: for open and closed eyes.

The fitting process is represented in Figure 6.3. We firstly match the global AAM that gives us a rough eye modelling. Then, a blink detector is applied, determining if the eyes are open or closed; the blink detector is described in section 6.1.3.2. Next, the corresponding sub-model is applied, i.e., the closed-eyes sub-model if the detector indicated a blink. The local sub-model accurately models the eyes.

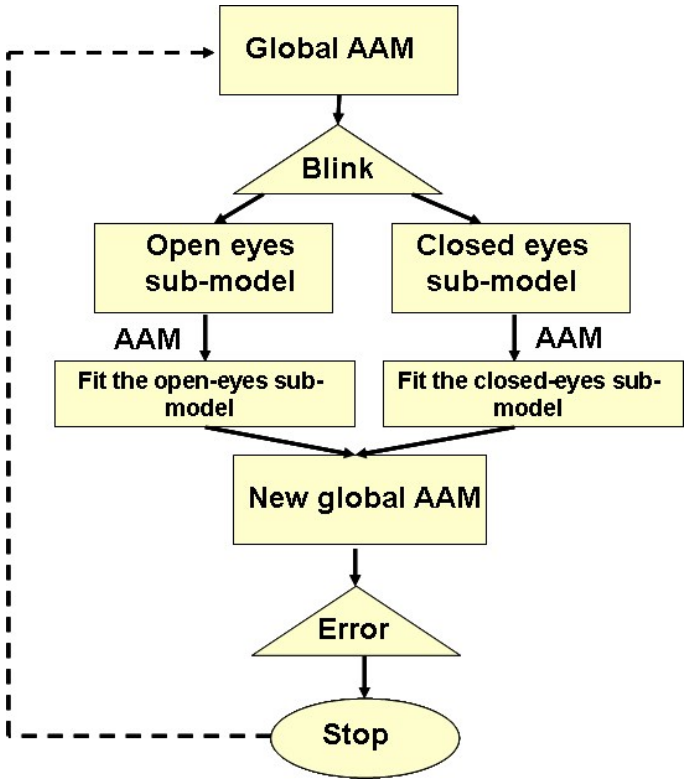


Figure 6.3. The fitting algorithm for the open/ closed eyes sub-models.

Only one of the open or closed sub-models is used in the fitting process, saving computational time. Another advantage is the accuracy of the shape annotation, as closed-eye shape is no longer obtained from open-eye shape. In consequence, fewer errors are introduced in the appearance model. The blinking information is still extracted thanks to the global model, but the accuracy of the shape is refined by the relevant sub-model.

Component-based eye AAM, independently modelling the left and the right eye

The idea of our second approach started from our idea to model each eye independently. A two-eye global model is necessary for an accurate location and initial modelling. Moreover we would like to be able to optimise the matching of the two eyes independently.

A global model is created using both eyes open and closed. Then a separate sub-model is created describing a single open or closed eye and the different variations in between the two states. The two models, i.e., one global model and one sub-model, are trained and are independently generated using the same AAM approach, as the one described in section 3.2.

One valuable aspect of the eye-model is the symmetry between the two eyes. This characteristic permits, starting from one eye model, e.g. left eye, to auto-generate an equivalent model for the other eye, e.g., right eye, respectively, simply by mirroring the data. The advantage is that we have less memory space requirements, as only one set of eye-model characteristics needs to be stored.

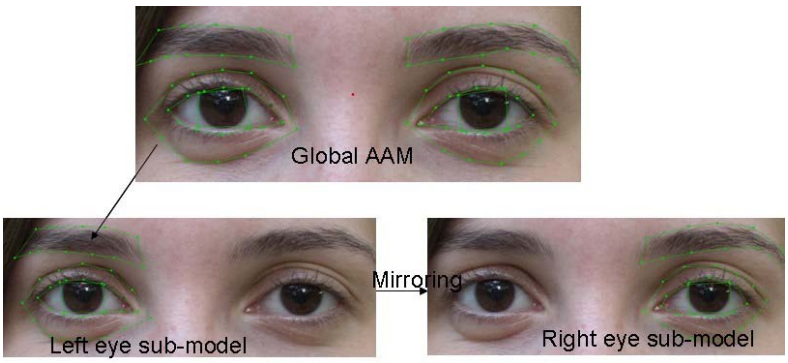


Figure 6.4. Examples of annotation for the global model, for the local sub-model and its mirroring in order to obtain the right eye.

The advantage of this version is that it permits that the two eyes find their optimal position and shape independently. There are situations, especially when dealing with large pose variations, plane rotations, occlusions, or strong facial expressions, when the 2-D projection of the eyes loses the property of global symmetry. An example of such a situation was presented in Figure 6.1, where we exemplified poor fitting of the global AAM.

Figure 6.5 describes the fitting algorithm adapted for this component-based version. At each iteration, the sub-model is inferred from the global model. Its optimum is detected by an AAM fitting procedure. Then the fitted sub-model is projected back into the global model.

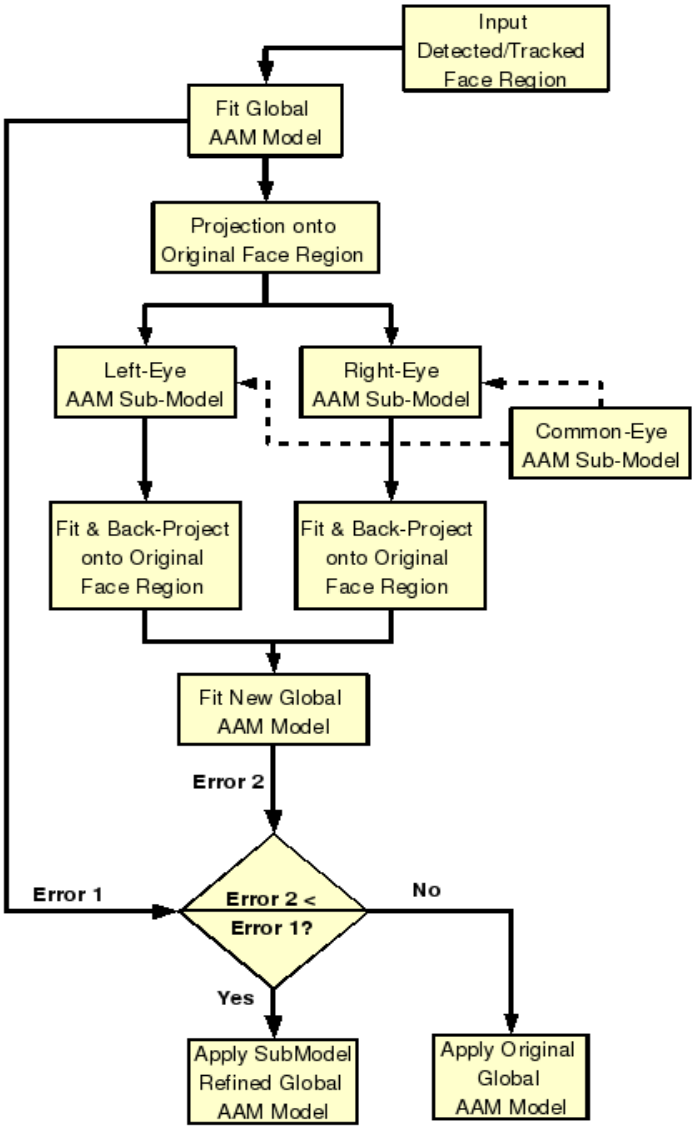


Figure 6.5. Fitting algorithm for component based AAM eye model.

In a first step, the global AAM is fitted, roughly locating the two eyes. The points of the global model which correspond to the sub-models are firstly extracted to form the two shape vectors, providing a good initialisation for each sub-model, which is fitted using the AAM method.

Next, we reintegrate the fitted points of the two independent sub-models back into the global model. Another projection of the global shape vector onto its principal components space is required. This step is necessary in order to constrain the two independently obtained eyes such that they remain within the limits specified by the global model.

In Figure 6.6, we present examples of fitting when we omit to reintegrate the fitted sub-models in the global model (Figure 6.6.b), compared to when we apply this

constraint (Figure 6.6.c). We also show the standard AAM result (Figure 6.6.a), for comparison.

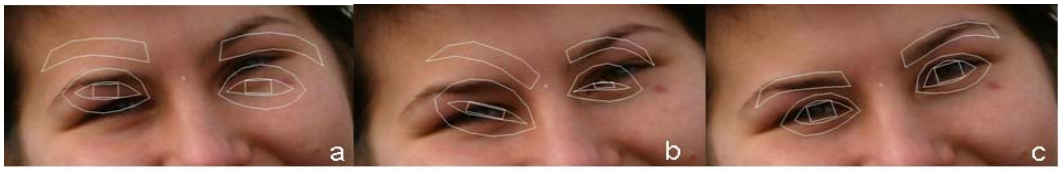


Figure 6.6. Fitting the standard AAM model; b. fitting the left/right eye sub-models without refitting the global model; c. fitting the left/right eye sub-models and refitting the global model.

In the last step of the fitting process, the fitting error for this refined global model is compared with the original global model fitting error and a decision is taken to use the global model with the least error.

6.1.2. Comparison of the Proposed Component-based Approaches

We proposed two different versions of adapting the component-based AAM for the eye region. The first version locally models both eyes, separating the open and closed eye situations. The second version independently models each eye, but it simultaneously includes open and closed eyes. We tested both versions, with a view to using the more effective model in a range of applications.

Both versions were trained using the same training set as for the standard eye model developed in section 5.1.3. 70 pictures were used presenting a high degree of variability in terms of subject individuality, pose, illumination, etc. Samples of the training set were already exemplified in Figure 5.6.

In order to compare the two versions, we used the same test set as the set used for testing the standard AAM eye model. We have described the test set in section 5.1.4. A detailed description of testing the component-based AAM compared with the standard AAM formulation will be given in Chapter 7.

For now, we conclude that the second version of the component-based AAM proved more useful and effective across a range of applications. In this version, we use as local sub-models single-eye models, adapted for open and closed eyes. This approach proved capable of accurately determining the eye parameters, particularly in cases of facial expression, where the two eyes deform independently. It is especially robustly to pose variations and to in-plane rotations.

The first version that we proposed, besides a poorer fitting, has the drawback that it is critically dependent on the accuracy of blink detection. This drawback is

explained in Figure 6.7. In the first picture the blink detector fails by indicating open eyes. This causes a catastrophic failure of the algorithm, as the open eye sub-model is chosen for fitting.

In the second version of the component-based AAM, each eye is fitted independently. Even if the global model fails by matching open eyes, the sub-models correctly match the eye image with closed eyes. This situation is represented in the second picture.

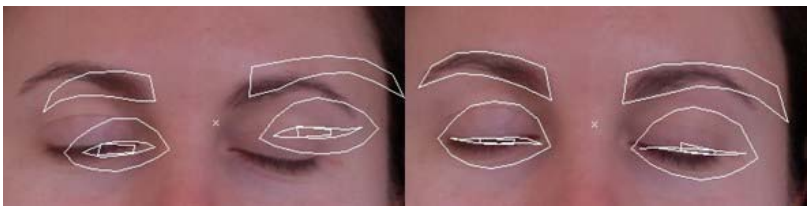


Figure 6.7. Comparison of the two proposed component-based versions: the two-eyes sub-model vs. the single-eye sub-model. It can be noticed that the first version fails as the blink detector detected open eyes.

The second version that we proposed, the component-based eye AAM independently modelling the left and right eye, will be employed in the applications below and in our work on expression recognition in Chapter 8.

6.1.3. Practical Uses of our Component-based AAM Eye Model

Extending the statistical eye-model of appearance to a component-based AAM permits the extraction of useful eye-related information. Shape and appearance parameters are determined; they describe the position, the degree of aperture of the eyes, and the gazing direction. This knowledge permits us to develop applications that track the eyes and detect a blink, or that are even capable of inferring facial expressions (see Chapters 8 & 9).

In the next sections we describe two applications: an eye tracker, in section 6.1.3.1, and a blink detector, in section 6.1.3.2. We use these two applications in order to test our eye model in various situations with unseen images.

6.1.3.1. Eye Tracker

Using the previously described eye model, we propose a robust and accurate algorithm to track the eyes. The robustness refers to pose variation, eye actions, e.g. blink, wink, gaze, and occlusions. This approach is summarised in Figure 6.8.

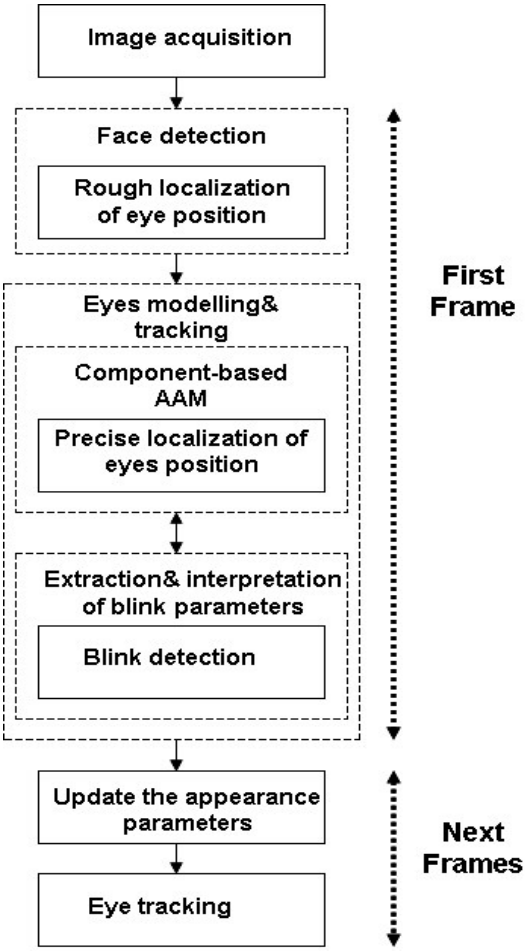


Figure 6.8. The eye tracker system.

Model Initialisation Scheme Based on Efficient Face Detector Estimates

Firstly, a face is detected in a frame of the video sequence, using the V-J face detector (see section 2.4.1). The initialisation step of the eyes is tuned with the V-J detector using a statistical relation learned from the training set. The statistical relation between the face detector estimates the face position and size, as a rectangle region, and the position and size of the reference shape inside the image frame is initially learnt offline from the training images. This relation is then used to obtain a more accurate initialisation for the reference shape, tuned with the V-J face detection algorithm. The approach of using this statistical relation is based on the idea proposed in [82].

It is important to have a reasonably close initialisation to the real values in order to insure the convergence of the fitting algorithm. Thus, initial rough estimations of the eye position, 2D rotation angle, and scale are deduced and are used as initial inputs for our eye-model.

Eye modelling stage

Then the model appearance parameters and, if required, blink information, are extracted. All the parameters are initially optimised by the component-based AAM procedure described in section 6.1.1. They include shape parameters associated with pose and blink, texture parameters, as well as parameters which characterise the modelled eye-patch inside the image frame. The purpose of this stage is to get an initial strong eye position from the first frame of the image sequence.

This step is the most computationally expensive. Preferably it is performed on the first frame and it is only periodically repeated. In our experiments, this step is performed every 10 to 30 frames, but the frequency of initialisation depends on the application, quality of input video sequence, resolution, etc.

Eye tracking

After parameter initialisation, the eye parameters can be tracked in all the following frames. An update on each new image frame is made only on the global model parameters that encode shape variation, position, 2D rotation, and scaling. Since the computational requirements are substantially reduced, this operation is now performed on each separate frame. A summary of the eye tracking results and some visual exemplification will be detailed in section 7.1.2.

From our experience, we conclude that the reduced computational costs of an AAM eye tracking scheme, associated with the robustness to blink and gaze variations, can meet the requirements for real-time applications in embedded imaging appliances.

6.1.3.2. Blink Detection

The advantage of modelling the eyes utilising AAM techniques is that we obtain a detailed description of the eye and not just its rough location. Using an educated choice of geometrical distances of the synthesised shape, or the shape parameters itself, a blink criterion can be defined. We now consider a number of potential criteria which are useful for detecting eye-blink and quantifying the degree of closure of an eye.

Blink detection based on geometrical distances

Firstly, we chose as a blink criterion geometrical details. Our blink detector compares the ratio between the width of the eye and its aperture with a predefined threshold. The width of the eyes corresponds to the individual characteristics of the eyes and it is constant, whether the eyes are open or closed. The aperture of the eyes varies with the eyes state, i.e., it is at a minimum when eyes are closed and a maximum when the eyes are wide open.

The relevant geometrical distances are described in Figure 6.9. The chosen distances are between the eleventh and the fifth point on the vertical and the eight and the second point on the horizontal, respecting the annotation portrayed in Figure 5.5. By a trial-and-error method, we fixed the threshold at 1/6. The method is accurate as long as the AAM fitting algorithm provides correct eye location.

Note that eye gaze must also be taken into consideration, by introducing gaze pictures into the training set. If gaze is not considered, we observed that the model will fail to differentiate between eyes with a wide gaze angle and closed eyes. A detailed description of tests performed to verify this application will be given in section 7.1.3.

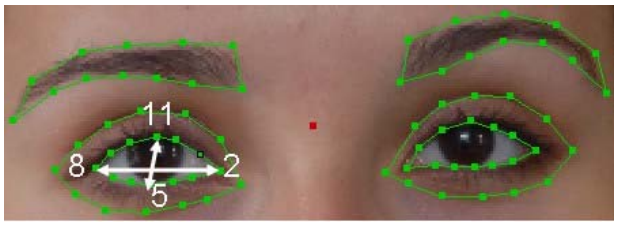


Figure 6.9. The proposed method for blink detection.

Blink detection based on shape parameters

In section 5.1.3.2 we have explained that blink shape parameters can be differentiated from the other types of shape parameters. In consequence, we have developed alternatives to the first proposed blink detector, such as identifying if the blink parameters belong to the open-eye curve, represented in green in Figure 5.7, or to the closed-eye one, the black curve in Figure 5.7.

The drawback of this method is that these curves are determined from a set of person-specific models, i.e., models based on training sets derived from single individuals. Also, the blink parameters have to be identified in advance. Thus the first method is preferable for its generality and simplicity. Summary of the blink detection results and their analysis will be given in section 7.1.3.

6.2. Advance AAM Lip Model Approach

In Chapter 5 we developed a lip model, based on the standard AAM formulation. We observed that the model failed our tests on unseen images in a proportion of 80% (section 5.2.3). We concluded that the main cause for this failure is a poor initialisation of the model. This is mainly due to the lip weak colour contrast of the lips and the significant overlap in colour features with the skin regions of the face (see Figure 6.10).

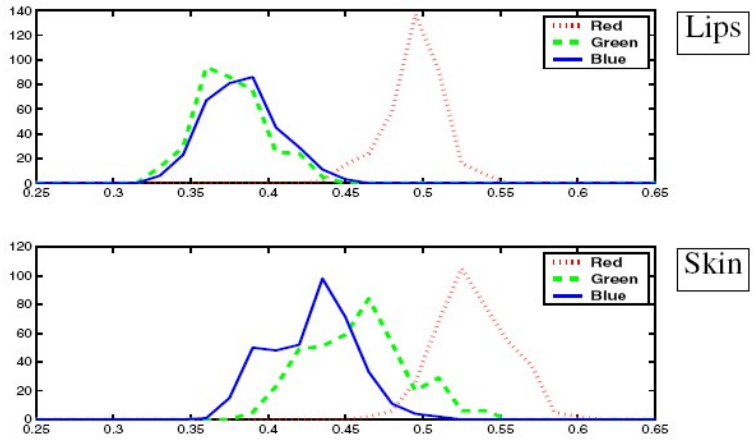


Figure 6.10. A practical exemplification of the weak contrast between lips and skin is described in [236] by typical R (dots), G (dashed) and B (plain) histograms of lips and skin; the histograms prove strongly similar for the two features.

In this section, we propose an improved version of our AAM lips model. We improve the standard AAM formulation by applying a pre-processing step that offers a more accurate initialisation of the lip region. In section 6.2.1, we describe this pre-processing initialisation step. The overall approach, embracing the initialisation and the AAM modelling, is described in section 6.2.2. We then test the performance of our lip model, by developing, in section 6.2.3, two consumer applications: a lip tracker and a smile detector. More thorough experimental testing of this model is documented in Chapter 7.

6.2.1. Initialisation of the Lip Region by Chrominance Analysis

The lips model requires a strong initialisation in order to achieve an accurate fitting to unseen images. Consequently we propose a pre-processing method that can provide such a robust initialisation. The most valuable information related to lips is their red colour, although red varies with respect to individuals, make-up, illumination etc.

Therefore, by filtering the red lip colour from the face region, we should be better able to identify the global shape of the lip region.

This approach is based on the work of Pantic et al [29] and it is adapted for our AAM models. Firstly, the input image is transformed into the HSV colour space, as hue representation is less affected by variations in illumination [224]. This colour space brings invariance to shadows, shading, and highlights and it permits using only the hue component for segmentation.

Then the object of interest is filtered into the red domain, by applying the following hue filter [29]:

$$f(h) = \begin{cases} \dfrac{1 - (h - h_0)^2}{w^2} & |h - h_0| \le w \\ 0 & |h - h_0| > w \end{cases} \qquad (6.1),$$

where h is the shifted hue value of each pixel so that $ho=1/3$ for red colour. ho controls the positioning of the filter in the *hue* colour plane and w controls the colour range of the filter around its ho value.

As the colour for the lip region varies with respect to person identity, light conditions, make-up etc., the challenge is to find optimal parameters for the filter. Although an optimal solution would be an adaptive hue filter as in [29], the simplified solution adopted in our case is to find the optimal parameters for predefined conditions, e.g., for a specific database.

A statistical analysis was performed on our training set (described in Figure 5.11). The variation of lip colour between individuals and the differences caused by varying illumination on lip pixels for the same picture was investigated. We noted that standard deviation does not vary much from picture to picture, as the pictures belong to the same database, with controlled conditions. The filter coefficients are chosen after performing an average on the mean and on the standard deviation for all pictures. The overall mean is 0.01 and it corresponds to the positioning of the filter h_0. The overall standard deviation approximating the filter width w is 0.007.

In our use of these techniques, after determining the parameters of the filter and performing the actual filtering operation, each image is binarised using a predefined threshold. The value of this threshold was set to 0.5, determined after a trial-and-error testing.

Morphological operations such as closing, followed by opening, can be used in order to fill in gaps and to eliminate pixels that do not belong to the lip region. After the lip region is determined, its centre of gravity (COG) is calculated. This point is used as the initialisation point for the AAM fitting algorithm.

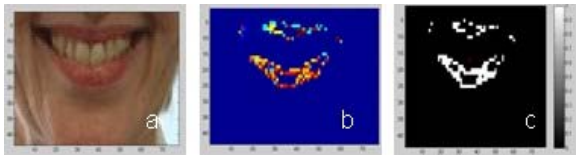


Figure 6.11. Lip region pre-processing: a. original image, b. after the hue filter, c. after the binarisation.

6.2.2. The Overall Formulation of the AAM Lip Model

Our lips modelling approach is composed of two main steps: an initialisation step and a modelling step.

Firstly, before the lip feature can be extracted and analysed, a face must be detected in an image and its features must be traced. The face is inferred from the V-J face detector applied on the input image. Then, our region of interest (ROI), i.e. the lip region, is deducted from the rectangle describing the surroundings of the face. ROI is reduced then to the lower third on the y axis, while 3/5 of the face box is retained on the x axis, as shown in Figure 6.12. A hue filter is then used to *provide an initial location of the lip region within this ROI.* This hue filter was defined in section 6.2.1.

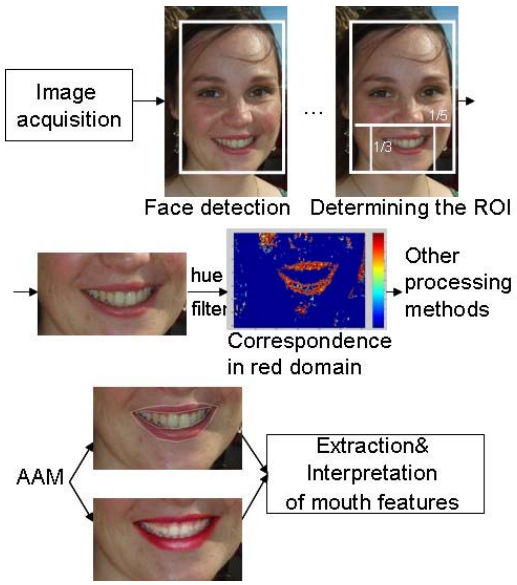


Figure 6.12. Lip modelling system overview.

In a second step, AAM is applied in order to perform a refined detection and to determine detailed lips features. The starting point for the algorithm is the COG of

the hue filtered ROI. The AAM adjusts the parameters so that a synthetic example is generated, which matches the image as closely as possible, as seen in Figure 6.12. Optimal texture and shape parameters are determined using the standard AAM lip model. In consequence information regarding lip features, such as its curvature, degree of opening, or its texture, can be determined.

6.2.3. Practical Uses of our Improved Lip Model

Our statistical lip model of appearance, improved with a pre-processing step, permits the extraction of useful lip-related information. Shape and appearance parameters are determined; they describe the position and the lip expression. This knowledge permits us to develop applications that track the lips and detect smiles, or even infer facial expressions (see Chapter 8).

In the next section we describe two applications which were developed from this research: a lips tracker, in section 6.2.3.1, and a smile detector, in section 6.2.3.2. We use these two applications to test our lip model with unseen images in a variety of situations.

6.2.3.1. Lip tracking

One application for the lip model and its corresponding fitting algorithm is a robust lip tracker.

In this section, we describe a framework for efficiently tracking the lips through an image sequence. Robustness to lip movements, slight head rotation, and head pose are our goals. A summary of the lip tracker results will be given in section 7.2.2.

The same principle as for the eye tracker described in section 6.1.3.1 is used to track the lips. A face detection stage is firstly employed in order to obtain an initial rough estimate of the face position, 2D rotation angle, and scale. These estimates are used to initialise the statistical model inside the image frame. Then, the lips model is initialised using a hue filter, as described in section 6.2.2. As our filter is not adaptive, it requires a predetermined parameter setting, following the procedure described in section 6.2.2.

The COG of the hue filtered ROI is next used as the initialisation point for the AAM algorithm. This step should be executed on the first frame and periodically

repeated. This step is the most computationally expensive and in the next frames only an update of the shape parameters is performed. Once this stage is performed, the model is able to deform and accurately track lip movements. As the computational requirements are now substantially reduced, this operation is performed on each separate frame. A detailed summary of test results for the lip tracker will be given in section 7.2.2.

6.2.3.2. Smile Detector

Another consumer application that can be developed using our model is a smile detector.

Smile detector based on geometrical distances

If an accurate analysis of the lip region is accomplished, it is possible to build a smile detector based on an educated choice of geometrical distances between lips. The procedure uses the same principle as for the blink detector described in section 6.1.3.2.

This type of information is provided by the shape parameters determined after the fitting stage. The width of the lip region can be used to determine if a smile has occurred. However it should be measured relative to a fixed distance and not relative to the width of the lips regions itself. This is because the width of the lips region can be highly variable. Examples of such distances are the face size, which is determined after the face detection stage, or the distance between the eyes, determined after modelling the eyes. The lip size is calculated between the two corners of the mouth, i.e., the 7th and the 1st points in Figure 5.12. The approach is described in Figure 6.13.

Figure 6.13. The lip width is calculated between the mouth corners and compared with the face width.

Smile detector based on a classifier

Another version is to use a simple classifier to classify between smile-pictures and non-smile pictures. This type of approach was employed in our experiments. We use a basic Euclidean-distance-based classifier. This classifier is trained using smile and non-smile pictures and uses the shape parameters. The results are presented in section 8.5.1.

Chapter 7 - Detailed Testing and Comparison of Modelling Approaches

In Chapter 6, we proposed extensions of the AAM models for the eye and lip regions. Here we summarise detailed testing and comparisons of our modelling approaches. We analyse their performance by comparing them with standard AAMs or by testing them in the context of several consumer applications.

Following the structure of the previous chapters, Chapter 7 is divided into two parts. In the first part, we test our component-based AAM eye model, while in the second part we present our lip model test results.

We start by analysing the performance of the component-based AAM eye model in section 7.1.1. We compare its results to the standard AAM formulation. Then, we analyse our eye model in terms of performance in various consumer applications, such as eye tracking, in section 7.1.2, and blink detection, in section 7.1.3.

In the second part of the chapter we summarise our lip model test results. We start in section 7.2.1 by testing the improvements brought by the hue filter initialisation. We follow its analysis, by testing it in various consumer applications, such as lip tracking, in section 7.2.2, and smile detection, in section 7.2.3.

We conclude our models performances and limitations in section 7.3.

7.1. Detailed Testing for the Component-based AAM Eye Model

The eye model is trained using the procedure described in section 5.1.3.2, using 70 pictures. As we want to develop a *general* model (section 3.5), we include a high degree of variability in the training pictures, in terms of subject individuality or illumination. Also, in order to model pose, blink, or gaze variations, we include these kinds of variation in our training set, based on the approach detailed in section 5.1.3.

The entire training set is displayed in Figure 7.1. It contains 30 open-eye pictures, 30 closed-eye pictures, and 10 pictures including small to medium head poses, with open and closed eyes. The subjects belong to both genders and various age categories.

Figure 7.1. The training set used to construct our eye model, containing 35 open-eye pictures and 35 closed-eye pictures. Between these pictures, there are eye images containing gaze and pose variation.

The pictures are *semi-automatically* annotated (see section 3.2.1) using 80 landmarks, distributed as described in section 5.1.3.1. An *independent AAM* model (see section 3.2.3) is then constructed. The resulting AAM model has 27 shape parameters and 39 texture parameters, as 95% of the shape and texture variances are allowed after PCA.

We tested our model using reduced allowed texture variances and at a resolution of 30*10 pixels. We observed that it still performs successfully with as little as 65% of variance. This corresponds to 21 shape and 5 texture parameters. The choice of the resolution and the allowed variance is motivated by the minimal resolution for which the model still performs successfully.

From our initial experiments we concluded that shape parameters contain more relevant information for our eye-related applications, e.g., from shape information we determine an eventual blink.

For our tests we use an adaptation of the Flexible Appearance Modelling environment (FAME) [237], modifying and extending it to accommodate the techniques described in this thesis. FAME is a software framework containing an open source implementation of the AAM. It was designed by Stegmann et al. and it is distributed as a public-domain material [237]. We want to express thus our gratitude to the authors of this laborious and comprehensive material for making it publicly available.

We tested our eye model on three test sets; none of the test images has been used in the training set, thus these tests are all on *unseen images*:

- Test Set 1: *Georgia Tech Face Database* [238] contains images of 50 persons, with 15 images per person. Most of the images are taken in two different sessions to take into account the variations in illumination conditions, facial expression, and appearance. In addition to this, the faces are captured at different scales and orientations.
- Test Set 2: *VidTIMIT Video Dataset* [239] is comprised of video recordings of different subjects reciting short sentences, but it is also useful for blink or gaze detection.
- Test Set 3: a *personal collection of the consumer facial images* containing 200 consumer pictures where the subjects blink, gaze, present head rotations, etc. The set is described in Appendix B.

7.1.1. Eye Model Comparisons: Standard vs. Component AAM Modelling

Using the model and test settings presented in the previous section, comparative results for the fitting accuracy between the standard AAM eye model and the proposed component-based AAM eye model are evaluated.

The V-J face detector described in section 2.4.1 is initially applied to localise the faces for the three test sets. The detector successfully locates the face region in 98% of our test images.

Qualitative testing

Firstly, we realise a qualitative testing by visual inspection. We observed that our improved model, described in Chapter 6, brings considerable improvements over our standard statistical model, described in Chapter 5, especially in cases of posing and in-plane rotations.

Both the classical and the component-based AAM work well for men, women, or children, and for various eye colours. However, the latter model shows a better fitting in case of under-trained models or in more complex situations that deal with:

- ***Pose****:* e.g., in head rotations, the left eye will have a slightly different shape from the right one which is captured by the component AAM approach.
- ***Occlusion****:* using the left/right eye sub-models, the eyes will fit independently and theoretically the non-occluded eye will present correct fitting, information that can be combined with additional image processing to fit the occluded eye.
- ***Face expression****:* one eye might change its shape independently of the other, e.g., winking.
- ***Unseen details of the eye****:* not seen in the training set, e.g., unseen shapes or distances between the eyes, distances between the eyebrow and the eye.
- ***In-plane-rotation****:* as represented in Figure 7.2.

An example of improvements of the eye region fitting when using the component-based method is presented in Figure 7.2.

Figure 7.2. From left to right: a. the initialisation of the eye-shape inferred from face detection; b. the eye-shape from global AAM fitting; c. the eye-shape from AAM sub-model fitting; d. the fitted shapes of the component sub-models refitted by the global AAM – note how the two eyes are constrained into a two-eye limit.

An important qualitative improvement is noticed when using the component-based AAM for winking eyes, even when wink pictures are not present in the training

set. This is possible because the component-based AAM allows the eyes to be modelled separately, so one eye can find its closed-eye shape, while the other eye is modelled as being open. This case is presented in the last row of Figure 7.3. The same figure gives other examples of better fitting in the case of the component-based AAM eye model.

Figure 7.3. Component based AAM vs. standard AAM for the eye region. In the first column, the standard AAM is fitted. The second column presents examples of the component-based AAM when the final constraint was not yet applied (the two eyes are fitted independently). The last column shows the component-based AAM results, when a final two-eyes-together constraint is applied.

Quantitative testing

After visually inspecting the results, a quantitative evaluation of the proposed model is performed on representative examples of the three test sets. The quantitative evaluation of a model performance is realised in terms of boundary errors, calculated as ***Point-to-Point (Pt-Pt) shape error***, calculated as the Euclidean distance between the ground-truth shape vector and the converged shape vector inside the image frame, as shown in equation (7.1):

$$Pt - Pt = \frac{1}{n}\sum_{i=1}^{n}((x_i - x_i^g)^2 + (y_i - y_i^g)^2)^{1/2} \qquad (7.1)$$

where the index g marks the ground truth data, obtained in our case by hand annotation.

Another type of error can be calculated, namely the ***Point-to-Curve shape error***. It is calculated as the Euclidian norm of the vector of distances from each

landmark point of the exact shape to the closest point on the associated border of the optimised model shape in the image frame, using equation (7.2):

$$Pt-Crv = \frac{1}{n}\sum_{i=1}^{n} min((x_i - r_x^g(t))^2 + (y_i - r_y^g(t))^2)^{1/2} \qquad (7.2)$$

where the border is modelled as a linear spline $r(t) = (r_x(t); r_y(t))$.

The mean and standard deviation of *Pt-Pt* and *Pt-Crv* are used to evaluate the boundary errors over a whole set of images. Figure 7.4 presents the histogram of Pt-Pt shape errors calculated with respect to manual ground-truth annotations. The ground-truth annotations represent hand-annotated shapes. In these tests we compare the standard AAM formulation, the component-based method, when omitting its last stage, i.e., the global fitting, and the component-based AAM with global fitting. The initialisation provided from the face detection step is used as benchmark.

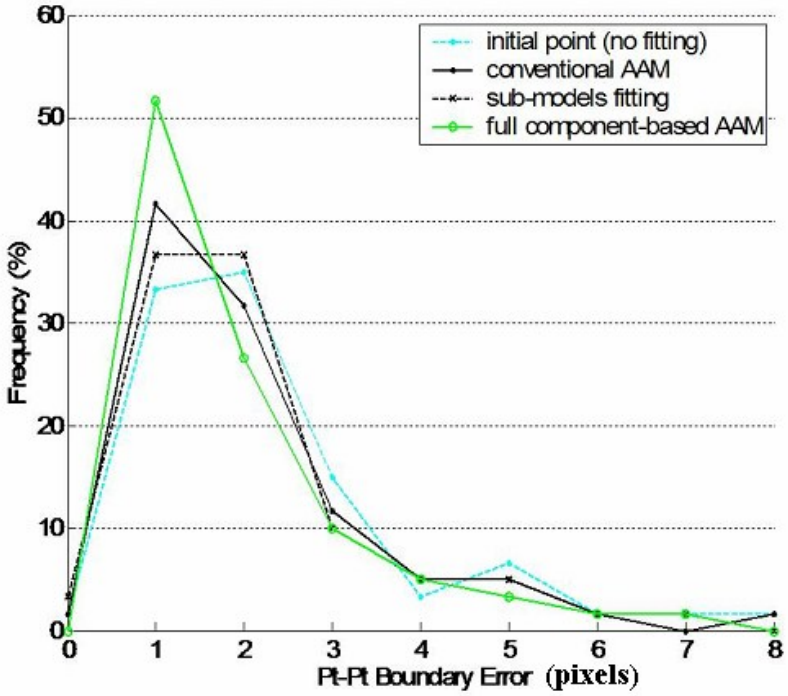


Figure 7.4. The histogram of the boundary error for the three algorithms: standard AAM, sub-model fitting, and the component-based AAM.

From the figure it can be observed that the boundary errors for the tested fitting algorithms are concentrated within lower values as compared to the initial point generated from the detection algorithm, showing the methods improvement for eye location. Furthermore, it can be noticed that the shape boundary errors are

concentrated within lowest values, indicating that the full component-based AAM performs the best in terms of fitting accuracy, thus resulting in a clear improvement over the initial position, as well as over the other fitting algorithms.

The presented results show the superiority of the proposed optimisation scheme in terms of lower boundary errors. More results for the component-based AAM can be evaluated while testing the eye tracker and the blink detector, in the next sections (Tables 7.1-7.3).

7.1.2. Testing the Eye Tracker Application

We next evaluate the eye tracker proposed and described in section 6.1.3.1 using 12 test video sequences on unseen images as follows: (i) three sequences from the Georgia Tech Face database [238], (ii) three sequences from the VidTIMIT Video dataset [239], and (iii) six video sequences collected especially for this experiment.

These video sequences are used to measure the performance during tracking. Our collection of video sequences contains rotations of the head, blinking of the eyes, and partial occlusions. Only face images where the V-J face detector has successfully reported a face have been used for these tests.

We used the trained models detailed in section 7.1. A specific-person model, specialised on each tracked subject, would represent an easier case scenario and it is imaginable that it would perform better. However, we chose to use our general AAM model, as we want to develop an application which can be generalised to a range of individuals and image acquisition conditions. Moreover, manually annotating images for each personalised model for each individual in the test sequence is very time consuming.

It is worth remarking however, that as a consequence of some poor results in preliminary tests, we have subsequently introduced some representative pictures, five for each dataset, from the Georgia Tech Face and VidTIMIT Video databases into our initial training set. Our training set contains now 80 images. We note that even introducing a small number of training images from each data set has greatly improved the model fitting to *unseen images* from those respective datasets.

Qualitative testing

Figure 7.5 shows examples of eye tracking for two video sequences from our personal collection containing pose variation, blink, and gaze.

Figure 7.5. Eye tracking in two examples of video sequences from our collection of video sequences using our model, when blinking, facial expression, and small pose variations are involved.

Quantitative testing

Table 7.1 shows the tested performance of the proposed system. In its first row, Table 7.1 firstly gives a summary of the performance obtained by our standard statistical model of appearance in terms of eye tracking accuracies on the three test sets. Then, in its second row, it reports the improvements brought by the component-based AAM on the same test sets. It is clear that the component-based model improves overall accuracy for the system for all the databases used in the test set apart from VidTIMIT where both models work equally well.

The results were collected in a trial with 15 observations per video sequence, where 10 of these observations correspond to the first 10 frames and the remainder correspond to randomly chosen frames. These frames were compared by a *Pt-Crv error* (defined in section 7.1.1) with the same hand annotated frames, considered as being the ground truth.

TABLE 7.1. SUMMARY OF THE SYSTEM ACCURACY (%) FOR THE STANDARD MODEL OF APPEARANCE.

Method	Georgia Tech Face Database	VidTIMIT Dataset	Our database
Standard AAM	92	98	75
Component-based AAM	97.5	98	79.24

The tracking system is automatically initialised by the component-based AAM procedure and it achieves an average accuracy of 91.5%. Tracking performances directly depend on the eye fitting in the first frame. Weak points can be mostly noticed when the face becomes partially occluded, e.g. by moving a hand in front of the face, or large horizontal or vertical poses, e.g., more than 30 degrees.

7.1.3. Results of the Blink Detector

The benefits drawn from our eye modelling method are analysed in the context of blink and gaze detection. We evaluate the blink detector based on geometrical distances as described in section 6.1.3.2. The iris, indicating gaze movements, and eye contours, indicating blink are analysed in order to test our eye model performance.

Tables 7.2-7.3 summarise the blink detection and gaze fitting accuracies on the three tests sets described in section 7.1.1 for our standard AAM eye model and for our component-based approach. The blink detector returns a flag that has the value 1 if a blink is detected and 0 otherwise.

The gaze detection rates were calculated on open eyes pictures with respect to the *Pt-Crv shape errors*, compared to ground truth annotations. 10 frames of each video sequence were evaluated.

TABLE 7.2. SUMMARY OF THE SYSTEM ACCURACY (%) FOR THE STANDARD MODEL OF APPEARANCE.

Method	Georgia Tech Face Database	VidTIMIT Dataset	Our database
Gaze accuracy fitting	68.75	52.08	40.25
Blink detection	90.625	100	67.92

TABLE 7.3. SUMMARY OF SYSTEM ACCURACY (%) FOR THE COMPONENT-BASED MODEL.

Method	Georgia Tech Face Database	VidTIMIT Dataset	Our database
Gaze accuracy fitting	78.125	83.33	69.18
Blink detection	96.875	100	79.24

The results proved to be encouraging, namely the component model is able to correctly analyse the eye, a fact that infers a precise blink detection. A drawback of the model and its algorithm is the accuracy to gazing, where future improvements are necessary.

We note that failure of the tracking algorithm is often due to unfocused images, large face offsets between frames, and partial occlusions, and, in particular, large gaze variations. The poorest results were obtained on our own collection of pictures, explainable by the fact that our images are more challenging in terms of facial variation that either of the standard databases used in the experiment. We also note that the Georgia Tech database presents pose variations, while the pictures that we used from VidTIMIT incorporate only frontal faces. This fact explains the high blink detection rates obtained in the case of the latter test database.

7.2. Detailed Testing for the Advanced AAM Lip Model

We trained our lip model as described in section 5.2.2. A high degree of variability is considered for the training pictures, in terms of subject individuality, illumination, head pose, and lip expression, as we want to develop a *general* model (section 3.4).

The training set is displayed in Figure 7.6. It contains 30 pictures, among which six present pose variation. The subjects belong to both genders, to various age categories, and present various facial expressions.

Figure 7.6. The training set used to construct our lip model, containing 30 pictures.

The pictures are *semi-automatically* annotated (see section 3.2.1) using 24 landmarks, distributed as described in section 5.2.2. We construct an *independent AAM* model (see section 3.2.3). The resulting AAM model has 12 shape parameters and 17 texture parameters, as 95% of the shape and texture variances are allowed after PCA. In our experiments we use the FAME environment [237], modifying and extending it to accommodate the techniques described in this section.

The proposed algorithm is tested on several specialised research databases and on our custom dataset. We firstly compared the standard AAM formulation and our proposed component model approach on the FERET database [165, 166] and on our custom dataset. Our dataset comprises 80 general images, in which subjects presented different facial expression, various poses, make-up, illumination conditions, image resolutions, etc.

We also tested the accuracy of our lips tracker scheme on the audiovisual speech VidTIMIT Video Dataset [239] which is comprised of video recordings of different subjects reciting short sentences, suitable for tracking. The smile detector was tested on the standard databases MMI [167], and FEEDTUM [169]. These databases display different subjects picturing facial expressions.

7.2.1. Lips Model Comparisons: Standard vs. Hue Pre-filter AAM

Using the models and the test settings presented in the previous section, i.e., the FERET database and our custom, comparative results for the fitting accuracy between the standard AAM lip model and the AAM lip model using a hue filter initialisation are evaluated.

The V-J face detector, described in section 2.4.1, is initially applied to localise faces for the test sets.

Qualitative testing

Firstly, we realise a qualitative testing by visual inspection. A first observation is that although, ideally the hue domain should not be influenced by variations in illumination, this is not observed in practice. Therefore, changes of illumination can cause important variations in the hue component for the same object. Standard face databases are normally created using a controlled environment and thus the hue

components of all the images are much more consistent regarding differences between individuals.

In the general case of unknown illumination conditions, the parameters of the hue filter have to be adapted for each particular image, which is a very complex task. Yet, for a controlled environment, e.g. working with a standard database, the filter parameters, i.e. the mean and the standard deviation of the hue component of the lips, can now be estimated from within that particular tested database.

A statistical analysis is performed in order to investigate each database, from the point of view of the effect of the acquisition system over the texture, illumination variations, etc. of the pictures, as explained in section 6.2.1. Following this analysis, optimal parameters of the hue filter for images from that database can be readily determined.

We obtained the following parameters, summarised in Table 7.4:

TABLE 7.4. CALCULATED HUE FILTER PARAMETERS FOR EACH TEST SET.

Database	ho	w
FERET	0.03	0.005
MMI	0.01	0.005
FEED	0.01	0.004
VidTIMIT	0.02	0.004
Our collection	0.01	0.007

Examples of the results of the proposed fitting algorithm obtained for different pictures from FERET database are presented in Figure 7.7. The shape and texture parameters after the fitting stage are depicted. The first two columns represent the results of the AAM algorithm (respectively shape & texture), used as a stand-alone method, while the last two columns are a consequence of the hue filter combined with AAM. Note the improved alignment of the shape model and more natural texture of the improved hue-filter & AAM.

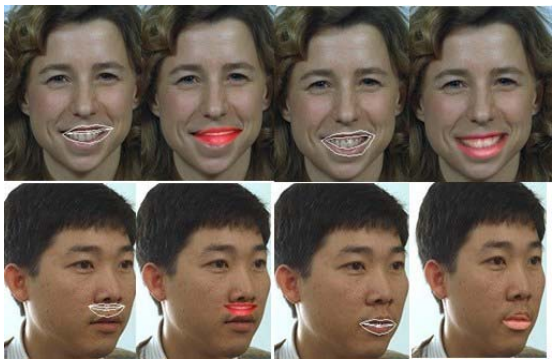


Figure 7.7. Mouth fitting results applying the AAM algorithm.

Quantitative testing

TABLE 7.5. FITTING ACCURACIES (%) OF OUR LIPS MODEL AS COMPARED TO A STANDARD AAM LIPS MODEL.

Database	Standard AAM	Our model
FERET	38	77
Our Collection	20	54

The results were calculated using the *Pt-Crv shape error* (see equation (7.2)), as compared to hand annotated images. The results showed the benefits of an accurate initialisation point for the AAM method.

The low rates can be explained by the fact that the filter parameters are not optimal for each picture, as they were calculated based on a statistical analysis on the picture collection. However, the pictures present a high degree of variability between them, so the statistical parameters might not be the most favourable for each picture. Improving the hue filter to adapt to each individual image is likely to boost these performance figures significantly but we did not have time to investigate this approach further.

7.2.2. Results of the Lip Tracker

The benefits drawn from our AAM enhancement are analysed in the context of lip tracking. For the tracking experiments, short movies of sequences of subjects smiling or reciting sentences are used. The test sets are selected from the VidTIMIT database and from our test sequences.

The general model developed in section 7.2 is used, and not a person-specific AAM model. However, some representative pictures from the test sets were included in the training set.

Our algorithm is applied for the first frame and refreshed periodically. Applying the technique on the VidTIMIT database and on our test sequences, it shows good tracking results for all individuals up to 30 degrees of face orientation. Sudden movements of the head affected the tracking performances. In this situation, the fitting algorithm had to be applied more often in the video sequence. In our case, the initialisation was applied every 10 frames. An example of the results obtained with our mouth tracker is presented in Figure 7.8. The summary of results is presented in Table 7.6.

Figure 7.8. Example of mouth tracking sequences: the first row contains results from one of our video sequences, while in the second row there are exemplified frames from the VidTIMIT database.

TABLE 7.6. SUMMARY OF THE SYSTEM ACCURACY (%) FOR THE LIP TRACKER.

Method	VidTIMIT (%)	Our collection (%)
AAM initialised from V-J	45	32
AAM initialised from the hue filter	78	67

7.2.3. Results of the Smile Detector

The benefits drawn from our AAM enhancement are analysed in the context of smile detection.

A smile detector is implemented based on a simple Euclidean distance, as defined in section 6.2.3.2. Its accuracy is verified on smiley and non-smiley pictures from the FEED and MMI databases.

TABLE 7.7. SUMMARY OF THE SYSTEM ACCURACY (%) FOR THE SMILE DETECTOR.

Method	MMI (%)	FEEDTUM (%)
AAM initialised from V-J	60.71	50
AAM initialised from the hue filter	73	60

Another version of smile detection is tested, based on geometrical distances between the inner corners of the eyes and the outer corners of the mouth, as described in section 6.2.3.2. While the first distance is considered constant despite any facial expression, the second distance increases in the case of a smile. For this approach, the following results are obtained: 61% for FEEDTUM database, 78% for the MMI database, and 60.5% for our collection of consumer images.

7.3. Chapter Summary and Conclusions

In this chapter we summarised the results obtained by our facial feature models, described in Chapter 6. We analysed our extensions, by comparing them to the results obtained in the case of the standard AAM formulation.

Eye model conclusions

In the first part of this chapter, the performances of our component-based AAM adapted for the eye region are evaluated. We evaluated accurate fitting, eye tracking performances, and blink detection rates. Better results were obtained for all tasks in the case of the component-based AAM.

While the standard AAM is based on a rough initialisation inferred from the face detection step, the component-based uses the AAM fitting as a starting point. Moreover, the fact that each eye is aligned independently permits that each sub-model finds its optimum separately. The optimal results are obtained by constraining the two eyes within a two-eyes-together limit, which represents a new AAM alignment of the result from the independent fitting of sub-models.

Lip model conclusions

In the second part of this chapter, our lip model based on the AAM formulation, together with a hue filter, is evaluated. We tested accurate lip fitting, lip tracking, and smile detection rates.

Better results were obtained in case of our proposed method, when we use a hue filter to determine a robust lip initialisation. However, our filter needs to be adapted for each set of conditions.

Weak points of the lip fitting were noticed when the hue filter fails to find the correct initialisation point. Situations like bearded people, very weak contrast between mouth and skin, strong illumination variation, or lip colour not covered by our statistical analysis, are examples in which the hue filter is suspected to fail. For these situations, the filter parameters had to be recalculated for its specific constraints.

In the next chapter, we are going to use our facial feature models to develop a robust complete face model. This face model is going to be used to accurately analyse facial expressions.

Chapter 8 - A Component-based AAM Representation for Facial Expression Modelling

The detailed analysis of facial expression can offer a wide range of new applications and open new areas of research, as seen in Chapter 1.

In this chapter we build on our earlier work to analyse the feasibility of integrating AAM techniques into an automatated facial expression analysis system. One of the key challenges of facial expression analysis is the extraction of relevant facial features. Our earlier work shows that AAM techniques are useful for analysis of both eye and mouth regions. Here we provide details of a comprehensive set of experiments on facial expression classification and recognition for still images and we propose a component-based AAM facial representation. We proceed to test the robustness of this framework to the deformations of typical unconstrained facial expressions.

We start in section 8.1 by providing an overview of the key components of an automated facial expression analysis system. We have implemented most of these already; here we place emphasis on the expression recognition module.

The challenges that a standard AAM has to overcome when analysing facial expressions are reviewed in section 8.2. These challenges motivate the experiments and tests presented later in this chapter.

In section 8.3, the most relevant AAM features are analysed in the context of facial expression analysis. Firstly we investigate which type of AAM feature, i.e., shape, texture, or appearance, contributes the most in modelling of the facial deformations caused by facial expressions.

In section 8.4, we investigate the optimal AAM parameters for modelling facial expressions and we discard the less relevant or redundant ones.

Next, in section 8.5, techniques for expression classification/ recognition on still images are presented. Expression classification techniques are dealt with in section 8.5.1, by comparing two classifiers: the Nearest Neighbour (NN) and the Support Vector Machine (SVM) and determining their optimal settings for expression classification. While many other classification techniques are known we restrict our comparative studies to these two approaches: SVM is considered one of the most effective techniques across a wide range of classification problems whereas the NN

method is one of the simplest. Restricting ourselves to these two techniques will give a good indication of the likely variations due to the classification process. In section 8.5.2 the same classifiers are adapted for a multi-class classification problem, such as expression recognition. We summarise the overall performance of these two classifiers in section 8.5.3.

Section 8.6 defines and compares the roles of global and local features in expression analysis and a component-based AAM approach to improve system performance is proposed. We begin in section 8.6.1 by analysing the importance of individual facial features, i.e., eyes and lips, in facial expressions analysis. We use the eye and lip AAM models developed in Chapter 6, determining through a series of experiments their individual usefulness in an automated expression classification system. Then, in section 8.6.2, we propose a robust component-based AAM face representation adapted for expression variations, focusing on the most significant eye and lips features.

We end this chapter in section 8.7, by analysing the robustness of our proposed facial representation and its performances in expression classification/ recognition.

8.1. Facial Expression Analysis System Overview

Facial expression analysis is a complex task involving multiple steps. The conventional facial expression analysis schematic was already described in Chapter 2 and depicted in Figure 2.1. We recall that generally, an expression analysis system is comprised of three main subsystems: (i) a face detection module, (ii) a feature extraction module, and (iii) a classification module which determines a similarity between the set of extracted features and a library of reference feature sets. Other filters or data pre-processing modules can be used between these main modules to improve the detection, feature extraction, or classification results.

Now, we aim to adapt this conventional schematic to our requirements. Ideally, our facial expression analysis system should be able to handle various illumination conditions, plane and out-of-plane head rotations, different image resolutions, etc. Also, as it is considered for consumer applications, the system should be automated and able to perform in real-time from a user perspective.

Our proposed system adapted for our requirements can be summarised as follows. In the first module, the V-J face detection algorithm [126] is applied to the current image. This face detection step not only determines whether there is a face in the input image, but also provides an accurate initialisation for applying more detailed AAMs to perform feature extraction. A statistical relation between the estimates of the position and the size of the detected face rectangle, and the position and the size of the reference shape inside the image frame is initially learned from a training set of images. This relation is further used as starting point for the AAM search. In our case, the detector successfully finds over 90% of the tested face images. The main cause of failing is large pose variations, i.e., over 45 degrees. Other causes include low image resolution (i.e., small faces in the image) or strong variations in illumination.

Next, facial features are extracted using AAM. Facial model parameters are calculated from the extracted facial area, using the component-based approach described in section 8.6.2.

The last step in an expression analysis system is expression classification/ recognition.

We remind the reader that in section 2.3.1 we discussed two modalities of describing emotions in computer vision: the Facial Action Coding System (FACS) and the six universal emotions. In the same section, we stated that in our work we are going to analyse the latter and we justified our choice. The choice of decoding the six universal emotions is based on the straightforwardness of the universal emotions compared to FACS. Also, the analysis of the universal emotions gives us the possibility to evaluate our work versus the work of other researchers who use the same framework.

We also discussed that in general, in order to determine an emotional state, additional factors should be taken into consideration: social and personal knowledge of the subject, their individual range of facial expression, etc. Since in our work, we consider only the facial expressions as a source for human emotions, the six universal emotions may also be referred to as the six universal facial expressions. And, in addition, there is the neutral face expression or null emotion which adds a de-facto seventh class.

Thus our system takes a multi-class decision between the six universal facial expressions and the neutral one. For comparison we will perform this step using two

distinct methods: a ***Nearest Neighbour*** (NN) approach and a ***Support Vector Machine*** (SVM) approach. Both approaches were previously described in section 4.3.

These last two steps will be detailed in sections 8.3-8.6. Experiments will be performed on specialised databases as well as on our own custom dataset of images. For our experiments, we have chosen the FEEDTUM [169, 170] and MMI [167] databases not only because of their popularity among researchers, but also because they contain spontaneous (FEEDTUM) and posed (MMI) facial expressions of various subjects of both genders and of all ages and ethnicities. These databases were described in greater detail in section 4.5.

Unfortunately, none of the publicly available databases presents a large range of facial deformations under a wide range of head poses and illumination conditions. Thus, our experiments will also be performed on our own custom dataset, which incorporates both pose and expression variations. We have built this test dataset containing eight subjects pictured in the six universal expressions and the neutral one, at different out-of-plane rotations. Some examples from our collection of pictures are shown in Figure 8.1. A detailed description of our image set is given in Appendix B.

Figure 8.1. Examples of consumer images containing different subjects, presenting both small to medium pose variations and facial expressions (left to right, happy, sad, surprised, and disgusted).

8.2. Motivation for our Research

In section 4.2.1 we have described limitations of the standard AAM formulation when modelling facial expressions:

- ***unsatisfactory AAM fitting rates on unseen facial images***, i.e., pictures not used in the training set, unfamiliar faces for the training set. In practice, in the fitting stage, faces can become too closely constrained to the model variations learned during the training phase, causing poor fitting for unseen facial images.
- ***poor initialisations*** of the model;

- ***lack of robustness to facial expressions deformations***, head poses, illumination variations, or occlusions;
- ***large number of AAM appearance parameters*** which may not correspond to the optimal parameters when decoding facial expressions.

Two practical examples of poor AAM fitting in cases of unseen facial expression and head pose are given in Figure 8.2.

Figure 8.2. Examples of poor AAM fitting on happy/neutral faces, with unseen variations, caused by facial expression, head pose, or variation in illumination.

Based on the above listing of issues with standard AAMs when applied to the problem of facial expressions it is clear that a more sophisticated approach needs to be taken to ensure better fitting of both the main facial region and the sub-regions, i.e., eyes, lips, and feature points which are most significant in a determination of the state of facial features and the corresponding facial expression.

Here we will shortly propose and examine an approach to improve the global and local fitting and the model registration for unseen faces and facial expressions. This will build on the earlier work described in Chapters 6 & 7.

But firstly, we search to optimise the speed of the system, by choosing the most relevant AAM model parameters/features. Our approach is described in the next section.

8.3. Relevant AAM Features for Illustrating Facial Expressions

Two types of features can be extracted when applying an AAM (see section 3.2): geometric features and features depicting the texture. If required, ***appearance*** parameters can be obtained by applying PCA on the concatenated geometric and texture parameters. ***Feature selection*** consists of keeping the most relevant facial features for classification and discarding irrelevant or redundant features. The quality of the extracted features plays a key role in their classification.

Both types of feature are important when fitting an AAM model, e.g., for the face, eye, or lips region, to unseen images.

A question we now need to ask is which of these parameters are more significant for expression decoding and which is the most appropriate approach to determine that?

In section 4.2.4, we have summarised the published research on this topic. We concluded that opinions are divided in the literature. Researchers agree that shape features have a large role to play in facial expression recognition [68, 97], but some are adamant that both shape and texture features are required [140] in order to obtain satisfactory results.

Let us begin with a qualitative analysis. On one hand, the skin texture of the facial region exhibits only slight changes during facial expressions. These minor changes arise due to local lighting effects, blushing, or wrinkling of the skin. Such changes can be considered relatively invariant when compared with the more obvious changes of shape observed for different facial expressions. Such geometrical features are very directly related to expressions. As examples, when surprised, we widely open the eyes and eventually the mouth, resulting in an elongated chin; when sad, we often blink; when angry, the eyebrows are usually drawn together.

So, we concur with the hypothesis that shape parameters are the most significant features in facial expression decoding. We now conduct some experiments to verify our initial hypothesis. In the experiments described in this section, we use a global standard AAM build for the face region.

Experimental Series 8.1- Choosing the optimal type of AAM feature

Goal: *find the most representative type of AAM feature for modelling facial expression*

Databases used: *FEEDTUM*

Number of training pictures: *30 for each classifier (15 represent the targeted facial expression, 15 comprise examples of the other categories of expressions)*

Number of tested pictures: *150 for each classifier (50 represent the tested expression and 100 random facial expressions, excluding the tested one). The pictures were not previously used in the training set, but can contain the same subjects as the ones in the training set.*

Classifier method used: *7 RBF- based SVMs with* $\delta = 2^2$ *(described in section 8.4.1.2) classifying between expression and non-expression pictures*

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters used for classification: *95% parameter allowed variance, resulting in 18 shape, 22 texture, and 19 appearance parameters*

For this experiment shape-only, texture-only, concatenated shape & texture, and merged shape & texture parameters are each tested. We are searching for the best compromise between a reduced number of parameters and high expression decoding accuracy.

In order to perform expression classification, we build seven SVM classifiers, for each universal expression and the neutral one (see section 8.4.1.2) of the type expression/non-expression, following the reasoning detailed in section 4.3. A comparison of our experimental results is presented in Table 8.1.

TABLE 8.1. CLASSIFICATION ACCURACIES, WHEN USING SVM, FOR SHAPE PARAMETERS, TEXTURE PARAMETERS, SHAPE AND TEXTURED PARAMETERS CONCATENATED INDEPENDENTLY, AND A JOINT COMBINATION OF THE TWO.

Classifier	Shape (%)	Texture (%)	Shape & Texture Independent (%)	Shape & Texture Combined (%)
Happy/Non-happy	92.66	80	97.33	86.66
Disgusted/Non-disgusted	69.33	64	67.33	67.33
Surprised/Non-surprised	72	64.66	82	71.33
Angry/Non-angry	90	76	45.33	79.33
Sad/Non-sad	68.66	45.33	66.66	67.33
Fear/Non-fear	72	74	78	77.33
Neutral/ Non-neutral	61.33	64	70	69.33
Average	75.14	66.86	72.37	74.10

Based on the results of these tests we determined that shape parameters did indeed prove to be overall the most valuable features for facial expression decoding accuracy. Shape results are comparable with the results obtained when applying a combined shape and texture model, i.e., when using the appearance features. However, as the number of shape parameters is somewhat less, the computational requirements both for feature extraction and to subsequently perform classification are also reduced. Thus shape parameters on their own have a demonstrable advantage over texture-only and both concatenated and merged shape + texture. These finding were also confirmed by the authors of [97].

It should be remarked that the accuracy rates of classification are not specifically addressed in this series of experiments. Improvements aiming to increase the accuracy of expression classifiers will be presented in the next section.

8.4. Relevant AAM Parameters for Illustrating Facial Expressions

While we have just shown that the shape parameters of an AAM face model contain the greatest "information density" for determining facial expression, there is still likely to be a significant amount of redundant information within this parameter set. Contra-wise, there may also be significant useful information contained in a small subset of the texture parameters. Ideally we would like to be able to further refine our use of AAM model parameters to achieve a more optimal set of parameters which are more closely tuned to the requirements of facial expression analysis.

Thus, in the next two experiments, we begin by investigating the particular shape parameters that best represent facial expressions. We note that in the experiments described in this section, we use a global standard AAM build for the face region.

Experimental Series 8.2- The representative shape parameters for modelling facial expressions

Goal: *determine those particular shape parameters most relevant for modelling facial expressions and discarding the irrelevant/ redundant ones*
Databases: *our own database, FEEDTUM*
Number of training pictures: *42 (21 from our database and 21 from FEEDTUM), 6 per expression*
Number of tested pictures: *150 (50 from our database and 100 from FEEDTUM)*
Classifier method used: *Euclidean-based NN for expression recognition*
AAM model*: global standard AAM for the face region*
Number and type of AAM parameters used for classification: *17 shape parameters*

Figure 8.3 represents the mean over the AAM shape parameters of each of the six universal expressions and the neutral one of the training set that include different individuals, poses, and facial expressions.

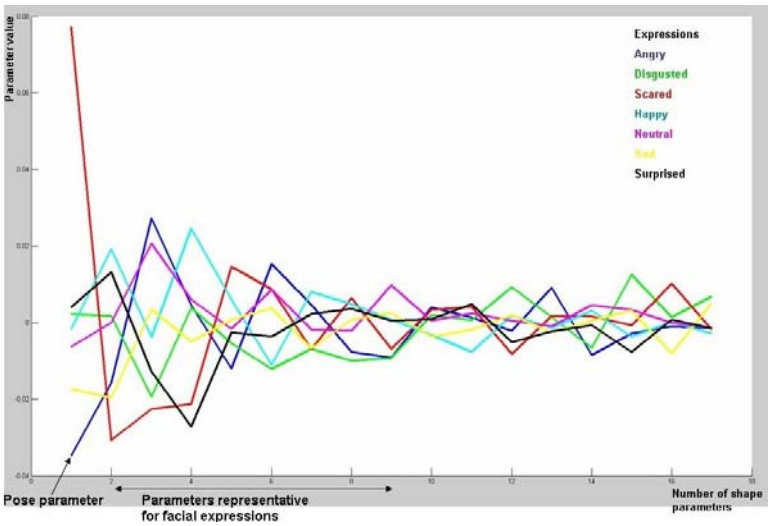


Figure 8.3. Mean over the shape parameters over each of the seven universal expressions.

Now, some shape parameters are likely to be redundant and others may reduce the system accuracy. In Figure 8.4, we have plotted the expression recognition rates versus the number of model parameters employed. A maximum of 17 shape parameters was used in total. Then, the number of parameters is reduced, eliminating in turn the parameter with the smallest variation across the test set of images. As the number of parameters is reduced we are left with the parameters which exhibit the widest variation.

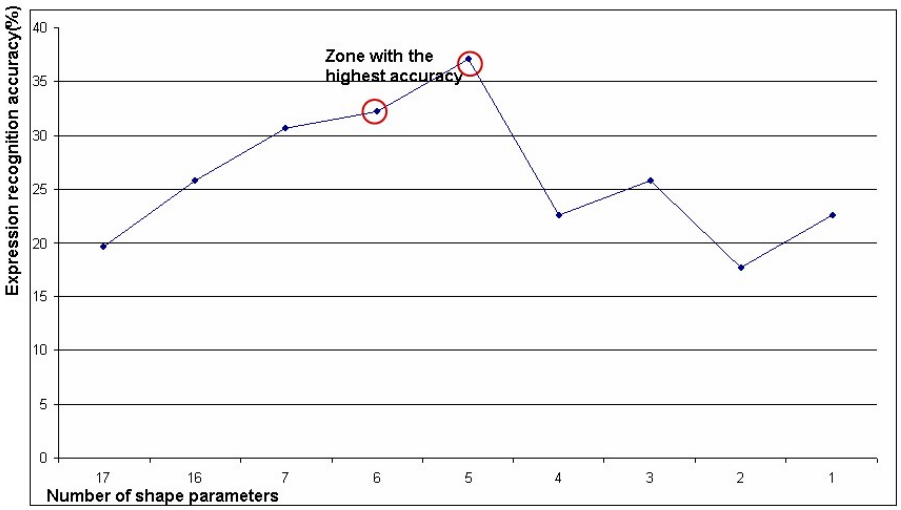


Figure 8.4. Expression recognition accuracies versus the number of shape parameters.

It can be noticed that optimal results are obtained when the five to six most variable parameters are used. As we increase the number of parameters beyond the 6^{th} parameter the model accuracy deteriorates.

In conclusion, an educated choice of parameters positively affects the system performances. In our case, it is proven empirically that using the first 30-40% of

model parameters provides higher expression recognition rates after the first parameter – which essentially corresponds to pose variation – is eliminated.

In this experiment we also noted that after performing PCA on shape parameters, the information on out-of-plane head pose was encoded in the first shape parameter. This is explained by the fact that the variations caused by pose cause significantly more geometric distortion than the variation caused by differences between individuals or between facial expressions. Consequently, pose variation is decorrelated to a large extent from other sources and manifests in the first-order PCA parameter.

The pose parameters are very important for the AAM fitting stage and the subsequent extraction of facial features. The pose variation information contributes to the shape and texture of facial features. However when analysing expressions, the pose parameters can be eliminated, as they do not contain the required facial expression information.

Experimental Series 8.3- Robustness to face pose

Goal: *test the robustness of our approach to face pose variations*
Databases: *our own database, FEEDTUM*
Number of training pictures: *42 (21 from our database and 21 from FEEDTUM), 6 per expression*
Number of tested pictures: *150 (50 from our database and 100 from FEEDTUM)*
Classifier method used: *Euclidean-based NN, RBF-SVM with* $\delta = 2^2$
AAM model*: global standard AAM for the face region*
Number and type of AAM parameters used for classification: *6 shape parameters*

In this experiment, we compare expression classification/ recognition rates for a set of shape parameters that include the pose parameter versus a set of shape parameters from which the pose parameter was eliminated. The results are summarised in Figure 8.5.

The differences between using/ eliminating the shape parameter are more significant in the case of expression recognition than in the case of expression classification. One possible reason is that in the case of expression classification, the classifiers are trained and tested just for two types of expressions, and not for seven as for expression recognition, though being able to better learn the pose variation. In consequence, the results are slightly comparable for using/ eliminating the pose

parameter. This is not the case for expression recognition, where the pose parameter decreases the recognition rates.

These experiments verify our hypothesis that the elimination of the head pose parameter helps decoding facial expressions.

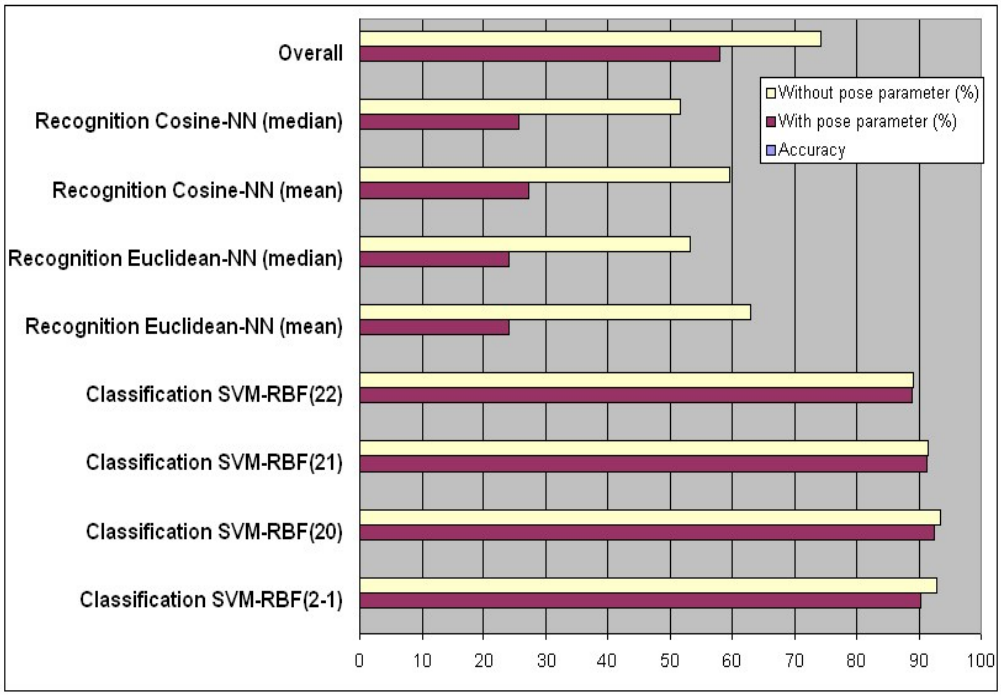


Figure 8.5. A comparison for the case when the pose parameters is taken into account or ignored. The benefits of eliminating the pose parameters are evident in all experiments, of classification or recognition, when using NN and SVM.

8.5. Expression Analysis on Still Images

This section presents a detailed investigation into expression analysis for still images. In the experiments described in this section, we use a global standard AAM build for the face region. In section 8.5.1 we test Nearest Neighbours (NN) and Support Vector Machines (SVM) techniques for expression classification. The same classifiers are also adapted and tested in the context of expression recognition, in section 8.5.2. We conclude the current section with a comparison of the performances of the NN and the SVM classification techniques in section 8.5.3, and chose the most appropriate for our expression analysis system.

8.5.1. Expression Classification

Remember that our primary research goal is to establish the validity of our approach incorporating both global and local AAM fittings. In our work, two classifiers are compared: NN (section 8.5.1.1) and SVM (section 8.5.1.2).

The choice of the two classifiers is based on their positive results obtained in the literature. In [156] it is stated and it is proved by experiments for gender classification and face recognition that SVM is typically among the top two classifiers, and the other top ranking classifier is one of the Euclidean-NN rule or the cosine-NN rule. As we designed them, in this work these classifiers use the relevant AAM parameters to choose between facial expressions.

8.5.1.1. Nearest Neighbour (NN)-based Classifier

The principle of a NN-based classifier used for expression classification was detailed in section 4.3. Now, we execute some experiments with this technique to determine its practical utility. In particular we investigate several different types of similarity metric and templates for NN. There are several similarity metrics that are commonly used with NN, including the Euclidean, cosine, and Mahalanobis distances, as seen in section 4.3. Also, two types of templates can be employed, based on calculating the mean or the median over the parameters. We test these settings in the following two experiments.

Experimental Series 8.4- The choice of the NN classification metric

Goal: *make a comparison between the benefits of Euclidean distance and cosine distance in the context of classifying facial expressions using a NN and choose the one providing the best results*
Databases: *FEEDTUM, MMI, our own database*
Number of training pictures: *30 for the FEEDTUM/MMI database, 15 for our own database*
Number of tested pictures: *150 for the FEEDTUM/MMI database, 40 for our own database*
Classifier method used: *Euclidean-based NN, cosine-based NN*
AAM model*: global standard AAM for the face region*
Number and type of AAM parameters used for classification: *7 shape parameters*

From our experiments on NN-based expression classification we determined that Euclidean and cosine distances perform almost equally well, but there is a slight advantage for the Euclidean distance from the point of view of a more consistent level of performance. Thus our future NN experiments will be based on the Euclidean

distance as being representative of the optimal NN technique. Some of the results are illustrated in Figure 8.5.

Experimental Series 8.5- Choice of the template used for comparison

Goal: *make a comparison between the benefits of mean or median techniques used to obtain templates for each expression; it will be chosen the technique giving the highest and most consistent classification rates*

Databases: *FEEDTUM, MMI*

Number of training pictures: *30 for each classifier*

Number of tested pictures: *1st series-the training set*

2nd series-150 for each classifier

Classifier method used: *Euclidean-based NN*

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters used for classification: *7 shape parameters*

Each test image is classified *by comparing its shape parameters with a template corresponding to each expression class*. In our case, we applied mean/median over each 15 pictures depicting each of the two expressions to be classified. We obtained two templates for the mean approach and two for the median approach.

The classifiers are of type *expression 1/expression 2*, i.e., neutral/non-neutral, sad/non-sad, angry/non-angry, disgusted/non-disgusted, fear/non-fear, surprised/non-surprised, and happy/non-happy, and of type *expression/non-expression,* e.g., happy/angry, happy/disgusted, disgusted/angry, etc. Considering that we have six expressions and the neutral one, 28 classifiers are obtained, seven for the former type and 21 for the latter.

Now, in case of each classifier, we calculate the distance between the shape parameter vector of each image test with the shape parameter vector of each template. The template yielding the minimum distance is selected.

In our experiments, at first, the AAM training and test sets coincide, i.e., there are no unseen images. This is so that we do not bias our results with poor AAM fittings. The Tables 8.2-8.5 summarise the classification rates. The correct classification rates must be read by looking in the tables at the line corresponding to class 1 and the column corresponding to class 2. For example, in Table 8.2, the correct discrimination between the disgusted expression and the happy one must be read at the value at line 2, column 4.

TABLE 8.2. EXPRESSION CLASSIFICATION ACCURACIES (%) FOR THE MMI DATABASE (TRAINING AND TEST SETS OVERLAP) WHEN USING A NN WITH A *MEAN* TEMPLATE RULE- AVERAGE OF EXPRESSION CLASSIFICATION 83.67 %.

	A	D	F	H	N	Sa	Su
A	80	75	85	90	75	80	87.5
D		87.5	85	95	92.5	95	95
F			57.5	85	67.5	72.5	82.5
H				87.5	85	95	95
N					80	70	85
Sa						85	87.5
Su							85

TABLE 8.3. EXPRESSION CLASSIFICATION ACCURACIES (%) FOR THE MMI DATABASE (TRAINING AND TEST SETS OVERLAP) WHEN USING A NN WITH A *MEDIAN* TEMPLATE RULE- AVERAGE OF EXPRESSION CLASSIFICATION 83.41%.

	A	D	F	H	N	Sa	Su
A	82.5	90	72.5	77.5	77.5	90	85
D		85	90	82.5	90	97.5	92.5
F			67.5	75	75	85	82.5
H				77.5	75	87.5	90
N					90	72.5	85
Sa						87.5	88
Su							85

TABLE 8.4. EXPRESSION CLASSIFICATION ACCURACIES (%) FOR THE FEEDTUM DATABASE (TRAINING AND TEST SETS OVERLAP) WHEN USING A NN WITH A WITH A *MEAN* TEMPLATE RULE- AVERAGE OF EXPRESSION CLASSIFICATION 93.58 %.

	A	D	F	H	N	Sa	Su
A	100	96.7	93.4	100	96.7	96.7	100
D		80	90	100	70	90	100
F			83.4	96.7	80	93.4	100
H				90	90	100	100
N					100	93.4	100
Sa						80	100
Su							100

TABLE 8.5. EXPRESSION CLASSIFICATION ACCURACIES (%) FOR THE FEEDTUM DATABASE (TRAINING AND TEST SETS OVERLAP) WHEN USING A NN WITH A *MEDIAN* TEMPLATE RULE- AVERAGE OF EXPRESSION CLASSIFICATION 89.79%.

	A	D	F	H	N	Sa	Su
A	100	63.4	93.4	100	96.7	90	100
D		73.4	86.7	96.7	70	93.4	100
F			73.4	96.6	66.7	90	93.4
H				90	90	96.7	100
N					83.4	86.7	100
Sa						83.4	100
Su							100

As can be noticed from the experimental results (Tables 8.2-8.5), a template obtained by averaging the AAM shape parameters outperforms, slightly, a template based on the median approach.

Generalisation of the AAM to unseen subjects is tested in the second part of this experiment with *leave-one-subject-out* tests, in which all images of the tested subject are excluded from training. We observe a decrease in the classification accuracies. As expected, a standard AAM performs poorly for the case of unseen pictures. Improvements will be proposed in later sections.

Tables 8.6 and 8.7 summarise the corresponding results of these tests. It can be noticed also that the classifier performs better on the FEEDTUM database. This is due to the fact that MMI presents more variations in illumination, a fact that affects the AAM fitting and, consequently, the precisions of facial feature extraction, particularly for unseen images.

TABLE 8.6. EXPRESSION CLASSIFICATION ACCURACIES (%) FOR THE MMI DATABASE (TRAINING AND TEST SETS DO NOT OVERLAP) WHEN USING A NN WITH A MEAN TEMPLATE RULE- AVERAGE OF EXPRESSION CLASSIFICATION 62.99 %.

	A	D	F	H	N	Sa	Su
A	50	57.5	65	72.5	65	60	60
D		72.5	60	67.5	70	82.5	75
F			60	55	42.5	47.5	77.5
H				65	60	55	80
N					66	65	52.5
Sa						65	65
Su							50

TABLE 8.7. EXPRESSION CLASSIFICATION ACCURACIES (%) FOR THE FEEDTUM DATABASE (TRAINING AND TEST SETS DO NOT OVERLAP) WHEN USING A NN WITH A MEAN TEMPLATE RULE- AVERAGE OF EXPRESSION CLASSIFICATION 66.94%.

	A	D	F	H	N	Sa	Su
A	83.33	50	63.4	80	50	63.4	76.7
D		60	70	66.7	56.7	53.4	73.4
F			56.7	83.4	56.7	80	63.4
H				70	73.4	53.4	80
N					60	56.7	83.4
Sa						60	86.7
Su							63.4

We conclude that a template obtained by averaging the AAM shape parameters outperforms, slightly, a template based on the median approach. Thus in the experiments that follow, a template obtained by averaging the shape parameters is used for NN.

8.5.1.2. Support Vector Machines (SVM)

A theoretical description of SVM was given in section 4.3. Now, in our experiments, we research the optimal settings for SVM when applied to expression classification.

In experiment 8.6 we search for the best kernel function, but also for the optimal settings for each function.

Two functions are investigated: the RBF and the polynomial function. Best grade for polynomials and, respectively, optimal δ values for RBF are searched.

The method of hyperplanes separation is also investigated. We recall that various methods for hyperplane separation can be used [159], i.e., Quadratic Programming, Sequential Minimal Optimisation, and Least-Squares.

Experimental Series 8.6- Choice of the kernel function for SVM

Goal: *compare expression classification results for RBF and polynomial kernel functions, investigating also best parameter settings and the optimal hyperplane separation method*

Databases: *FEEDTUM, MMI*

Number of training pictures: *30 for each classifier*

Number of tested pictures: *150 for each classifier*

Classifier method used: *RBF and polynomial SVMs*

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters used for classification: *7 shape parameters*

Two strategies can be considered for adjusting the classifier parameters to classify facial expressions. The approaches depend on the amount of available training data. If there is a considerable amount of data available, a simple validation process is sufficient. Otherwise, a V-fold Cross-Validation or a Grid search needs to be implemented [159]. In our experiments, we consider the former case.

In the first part of the experiment, seven SVM classifiers of type expression 1/ expression 2 are built to distinguish the six universal expressions and the neutral one. The polynomial kernel, with order from 1 to 6 and the RBF kernel with δ from 2^{-2} to 2^{6} results are investigated. The results for the MMI database are detailed in Table 8.8.

TABLE 8.8. EXPRESSION CLASSIFICATION USING SVM FOR THE MMI DATABASE, WHEN USING DIFFERENT KERNELS.

Classifier	Polynomial-grade						RBF						
	1	2	3	4	5	6	2^{-3}	2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}
H/non-H	88.7	78.7	82.7	65.4	59.4	51.4	78.7	82	82.7	82.7	90	92.7	92.7
D/non-D	56	46.7	46	46	41.4	40.7	63.4	58	60.7	55.4	70.7	79.4	62.7
Su/non-Su	74.7	52.7	51.4	50	46	44.7	40.7	46.7	61.4	71.4	72	71.4	68.7
A/non-A	51.4	47.4	47.4	46	53.4	53.4	71.4	71.4	40	72.7	66	70	48
Sa/non-Sa	74	58	54	47.4	45.4	43.4	66	64	61.4	62.7	69.4	68.7	70.7
F/non-F	78.7	72	71.4	64.7	60.7	54	69.4	69.4	70	70.7	72.7	72	72
N/non-N	62.7	56	59.4	54.7	50.7	45.5	38.7	42.7	58.7	62.7	61.4	79.4	59.4
Average	69.5	58.8	58.9	53.5	51	47.6	61.2	62.1	62.2	68.4	71.8	76.3	67.8

The optimal kernel function proved to be RBF, with δ fixed on small values. It is our preferred choice based on the accuracy and the consistency of its results. In the second part of this experiment, we sought to confirm the previous conclusions, this time by using 28 classifiers, as in experiment 8.4. The results (Tables 8.9-8.10) confirmed that RBF with δ fixed on small values achieves the highest classification rates.

TABLE 8.9. ACCURACY (%) OF EXPRESSION CLASSIFICATION ON FEEDTUM FOR THE AAM SHAPE PARAMETERS WHEN APPLYING A SVM CLASSIFIER FOR RBF 2^2, AVERAGE OF EXPRESSION CLASSIFICATION 69.43%.

	A	D	F	H	N	Sa	Su
A	70	63	70	66.7	70	70	73.4
D		79.4	63.4	70	73.4	53.4	70
F			72	66.7	56.7	80	53.4
H				92.7	73.4	60	66.7
N					79.4	53.4	80
Sa						68.7	76.7
Su							71.4

TABLE 8.10. ACCURACY (%) OF EXPRESSION CLASSIFICATION ON MMI FOR THE AAM SHAPE PARAMETERS WHEN APPLYING A SVM CLASSIFIER FOR RBF 2^2, AVERAGE OF EXPRESSION CLASSIFICATION 62.59%.

	A	D	F	H	N	Sa	Su
A	62.5	55	65	52.5	52.5	55	65
D		65	65	52.5	60	62.5	77.5
F			55	62.5	60	52.5	77.5
H				57.5	57.5	65	75
N					52.5	57.5	77.5
Sa						57.5	75
Su							80

Although not explicitly described, in our experiments, we also experimented with the method of hyperplanes separation. The following methods were tested: Quadratic Programming, by default in all the other experiments, Sequential Minimal Optimisation, and Least-Squares. The results showed that the method of hyperplanes separation does not significantly influence the results.

We conclude that the RBF kernel function with δ fixed on small values gave the best results. Our findings are confirmed also in the literature, where the RBF kernel is the most common function to be used in expression recognition with SVM [107, 150, 162]. The choice is also motivated by theory. RBF has fewer adjustable parameters than any other commonly used kernel and it is easier to compute [159].

8.5.2. Expression Recognition

Now, after we trained a classifier to discriminate between two facial expressions (section 8.4.1), we would like to be able to associate a human face with one of the six universal expressions or the neutral one, i.e., to perform expression recognition. We

will again compare NN and SVM, this time adapted for a multi-class decision framework.

We begin, in section 8.5.2.1 with a series of experiments investigating the performances of NN in expression recognition. Then, in section 8.5.2.2, we describe techniques to adapt the binary SVM to a multi-class problem, such as expression recognition. We experiment with some of these approaches, making an empirical determination of the most suitable one.

8.5.2.1. Nearest Neighbour

Experimental Series 8.7- NN performances in expression recognition

Goal: *verify the performance of NN in expression recognition*
Databases: *FEEDTUM, MMI, our own database*
Number of training pictures: *30 for MMI, FEEDTUM, 20 for our own database*
Number of tested pictures: *150 for MMI, FEEDTUM, 50 for our own database*
Classifier method used: *Euclidean-based NN*
AAM model*: global standard AAM for the face region*
Number and type of AAM parameters used for classification: *7 shape parameters*

Table 8.11 summarises expression recognition rates for the standard AAM facial representation. It can be noticed that the highest recognition rates are obtained on the MMI database. This is explained by the fact that MMI has a better image resolution and less variation in illumination that the other dataset we used. And as we would expect, the poorest results were obtained on our own database which contains the strongest pose and illumination variations.

TABLE 8.11. SUMMARY OF THE SYSTEM ACCURACIES (%) FOR RECOGNISING EXPRESSIONS WITH EUCLIDEAN-NN ON THE ENTIRE FACE.

Database	Recognition rate (%)
FEEDTUM	35.71
MMI	40.88
Our database	22.66
Overall	33.08

8.5.2.2. Multi-class SVM

We recall (section 4.4) that by their nature SVMs are binary classifiers. However, there exist strategies by which SVMs can be adapted to multi-class tasks, such as

One-Against-One (1A1), One-Against-All (1AA), Directed Acyclic Graph (DAG), and SVMs in cascade structures.

In the following experiments we only exemplify the 1A1 and the cascade structures. The choice is based on their simplicity and their good results presented in the literature [107, 164].

Experimental Series 8.8- Multi-class SVM in expression recognition

Goal: *compare the performances of two alternatives of multi-class SVM (1A1 and cascaded SVMs) in expression recognition*

Databases: *FEEDTUM, MMI*

Number of training pictures: *30 for each classifier*

Number of tested pictures: *150 for each classifier*

Classifier method used: *RBF-SVMs with* $\delta = 2^2$

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters used for classification: *7 shape parameters*

In the first part of this experiment, we employ the 1A1 approach. 21 classifiers are applied for each picture in our test-bench. A general score is calculated for each picture. The "recognised" facial expression is considered to be the one which obtains the highest score. As an example, to calculate the score for a "happy face" we apply: happy/fear, happy/sad, happy/surprised, happy/neutral, happy/angry, and happy/disgusted. Every time that a happy face is identified a counter is incremented that represents its score. The results are summarised in Table 8.12.

TABLE 8.12. EXPRESSION RECOGNITION ACCURACIES FOR 1AA-SVM ON THE FEEDTUM AND MMI DATABASES.

Emotion	FEEDTUM (%)	MMI (%)
Surprise	71	76.1
Fear	66.3	62.5
Happiness	65.4	62.5
Anger	64.8	56.8
Neutral	62.8	60.4
Sadness	61.5	59.7
Disgust	57.6	64
Overall	64.2	63.15

In the second part of the experiment, another alternative to extend the binary SVM to a multi-class SVM is investigated. A cascaded structure consisting of the

most effective six classifiers of the seven which classifying the six universal expressions and the neutral expression are used. An alternative would be to use the seven classifiers, so as to have an "error" or unclassified category for expressions that pass through the full cascade without being classified. However, in order to reduce the computational time, we choose the first alternative.

The workflow and the corresponding results are summarised in Figure 8.6. The figure shows the recognition rate after each stage of the cascade, e.g. our system correctly recognises a surprised face with a probability of 84.5% or an angry face with a probability of 70%.

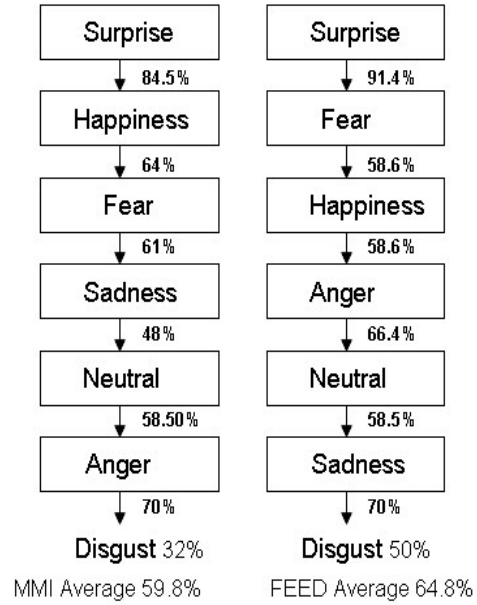


Figure 8.6. Performances of expression recognition of SVM classifiers in a cascade structure, for MMI and FEEDTUM databases.

8.5.3. Conclusion of Classifiers Performances

In section 8.5, we analysed approaches and corresponding results for expression classification and recognition in still images. Two classifiers, namely SVM and NN, were compared. After performing a series of five experiments we conclude the following:

- Overall, happiness proved to be the most recognisable expression, followed by surprise. These observations can be explained by the fact these particular expressions affect the shape of the facial features more than other expressions. Note, in particular, the open mouth and raised mouth corners. These expressions are followed by anger, sadness, disgust, neutral state, and fear.

- The results proved that the system is robust in dealing with subjects of both genders and several races and ages.
- SVM proved overall more effective as a classifier than NN, both from the point of view of higher classification/ recognition rates and the consistency of the results.
- Best results for SVM were obtained when using RBF kernel function with δ fixed on small values. These settings are going to be used in the next series of our experiments.
- Best results for NN were obtained when using the Euclidean distance and a template obtained from averaging the shape parameters as a metric for classification. These settings are also to be used in our next series of experiments.

8.6. Global vs. Local Features

Facial expressions are defined by the dynamics of individual facial features. Two types of representation can be used: global and local. The ***global*** representation considers the global properties of the pattern. For the ***local*** representation, a set of distinct geometrical facial features is determined. The importance of individual facial features is exemplified by the Thatcher illusion experiment [240], illustrated in Figure 8.7, that suggests that facial parts are processed to some extent independently by humans, and only loosely connected to the representation of the whole face.

Figure 8.7. The Thatcher illusion experiment [240].

Eyes and lips are considered the most distinct facial features in terms of depicting expressions. In section 8.6.1, we examine their independent contributions in the context of facial expression analysis in order to confirm our hypothesis.

In section 8.6.2, we make use of our individual feature models and of the global face representation in order to obtain a more robust component-based facial representation. We compare expression classification/ recognition rates of a global representation of the face and with rates obtained for our proposed component-based model and conclude on the best approach.

8.6.1. Relevant Facial Features for Illustrating Emotions

Eyes play an essential role in relaying feelings, emotions, and thoughts. People tend to look downward when sad or embarrassed. They tend to do the opposite, e.g., open widely the eyelids, when surprised. If the eyebrows are drawn together, it is often an indicator of an angry face.

Now, it is interesting to investigate how much information the eye-area brings to facial expression analysis. Would this area be sufficient to describe facial expressions? If completely and accurately modelled, how many details does it provide in order to recognise expressions? Is the lip area also relevant when analysing facial expressions?

To answer these questions expression classification and recognition for the eye and lip areas are now performed using the eye and lip AAM models developed in Chapter 6, adapting their training and test sets for expression analysis. A complete summary of these results is presented in the next two experiments.

Experimental Series 8.9- Expression analysis using the eye area

Goal: *investigate the quantity of facial expression information carried by the eye area and investigate its discriminatory power in classifying/recognising expressions*
Databases: *FEEDTUM, MMI*
Number of training pictures: *30 for each classifier*
Number of tested pictures: *150 for each classifier*
Classifier method used: *Euclidean-based NN classifier*
AAM model*: global standard AAM for the eye region compared to global standard AAM for the face region*
Number and type of AAM parameters used for classification: *11 shape parameters*

Results for expression recognition obtained by shape parameters of eye models, compared with the ones obtained for the whole face are summarised in Tables 8.13 and 8.14. On average, around 30% recognition accuracy is obtained performing a

simple NN classification on the entire set of shape parameters. This is a lower percent than the one obtained for the entire face using the same method in the same conditions. We note that the standard eye AAM described in Chapter 5 is used in our experiments and not the component-based AAM described in Chapter 6. The fact that the fitting accuracies are lower in the case of the standard AAM (Figure 7.4) affects the classification/ recognition rates. However, the performances of the standard AAM eye model are sufficient in order to make a suggestion of the importance of the eye feature in analysing facial expressions. The reasoning is also valid in case of the AAM lip model in the experiment 8.10.

Around 60% classification accuracy is obtained using a NN classifier for the eye area shape. The rates represent over 70% from the rates obtained for the entire face. Altogether, the figures indicate that a high amount of expression information is comprised by the eye shape.

Experimental Series 8.10- Expression analysis using the lip area

Goal: *investigate the quantity of facial expression information carried by the lip area and investigate its discriminatory power in classifying/recognising expressions*
Databases: *FEEDTUM, MMI*
Number of training pictures: *30 for each classifier*
Number of tested pictures: *150 for each classifier*
Classifier method used: *Euclidean-based NN classifiers*
AAM model*: global standard AAM for the lip region compared to global standard AAM for the face region*
Number and type of AAM parameters used for classification: *6 shape parameters*

Results for expression recognition obtained by shape parameters of lip models, compared with the ones obtained for the eye area and the whole face are summarised in Tables 8.13 and 8.14. Naturally, lips are also important emotion feature carriers, especially for expressions like happiness or surprise, with around 60% of independent contribution. However, when using a combination of eyes, lips, together with other facial information, the accuracy of an expression decoder increases as indicated by the classification/ recognition rates obtained on the entire face region.

TABLE 8.13. EXPRESSION RECOGNITION ACCURACIES WHEN ONLY THE EYE AND, RESPECTIVELY, THE LIP-AREAS ARE CONSIDERED.

Database	Recognition rate for the eye area (%)	Recognition rate for the lip area (%)	Recognition rate for the face area(%)
FEEDTUM	27.37	27.37	35.71
MMI	34.1	26.95	40.88
Our database	22.57	27.01	22.66
Overall	28.01	27.11	33.08

TABLE 8.14. EXPRESSION CLASSIFICATION ACCURACIES WHEN ONLY THE EYE AND, RESPECTIVELY, THE LIP-AREAS ARE CONSIDERED.

Classifier	FEEDTUM			MMI		
	Eye-area	Lip-area	Face-area	Eye-area	Lip-area	Face-area
Surprised/Disgusted (%)	44.37	55.87	85.83	73.12	76.66	85.62
Surprised/ Happy (%)	28.75	69.37	88.33	68.75	75	86.25
Happy/Sad (%)	41.87	46.25	68.33	72.5	50	73.12
Surprised/ Sad (%)	33.12	51.66	93.33	69.37	52.5	77
Overall (%)	37.02	55.78	83.95	70.93	63.54	80.49

8.6.2. Component-based AAM

The low expression classification/ recognition rates obtained in the previous experiments when using a global, i.e., standard, holistic AAM face model confirm the limitations of this approach, as listed in section 8.2.

The poor performance of the standard AAM when used for expression analysis can be understood because this model is based on a global approach and is thus more sensitive to changes in configuration than to changes in local features. Such a representation cannot be sufficiently flexible to permit adaptation to wider ranges of facial variability, such as the deformations which are present in the majority of distinctive facial expressions.

From a practical perspective we note that it is not possible to include all possible variations of shape and texture in the training set. The overall degrees of freedom inherent in the model are restricted by the number of model parameters. Or to put this another way, a model which did incorporate practically all possible facial variations in its training set would have a correspondingly large set of model parameters making it unwieldy and impractical. Thus any practical model we construct must necessarily restrict its potential variations. To achieve real

improvements in our classification it would seem that we need more than a single AAM model.

In this section, we propose to adapt the component-based AAM approach [94] for facial expression analysis. This approach benefits from both the generality of a global AAM and the local optimisations provided by its sub-models. This approach is designed to add degrees of freedom to the global model, by taking into account the individual variations of facial features.

The principle of the component-based algorithm was previously explained in Chapter 6, when applied to the eye area. Figure 8.8 explains its adaptation for the entire face region, using two sub-models: both an eye model and a lips model. These sub-models are based on the approaches originally proposed in Chapter 6.

Figure 8.8 indicates that a holistic AAM face model is created using the standard AAM, following the procedure detailed in section 3.3. Then separate sub-models are created for the eye and lips features, using the approach described in Chapter 6. At each iteration, the sub-models are inferred from the global model. Their optimums are detected by an AAM fitting procedure, based solely on the shape parameters. Then the fitted sub-models are projected back into the global model.

In Figure 8.9 we describe one practical example of the benefits of a component-based representation, on an image containing both pose and expression variations. This shows how the sub-models improve fitting of the global model to the entire face region, which in turn improves the alignment of each sub-model with the local features which are some important to accurate expression recognition.

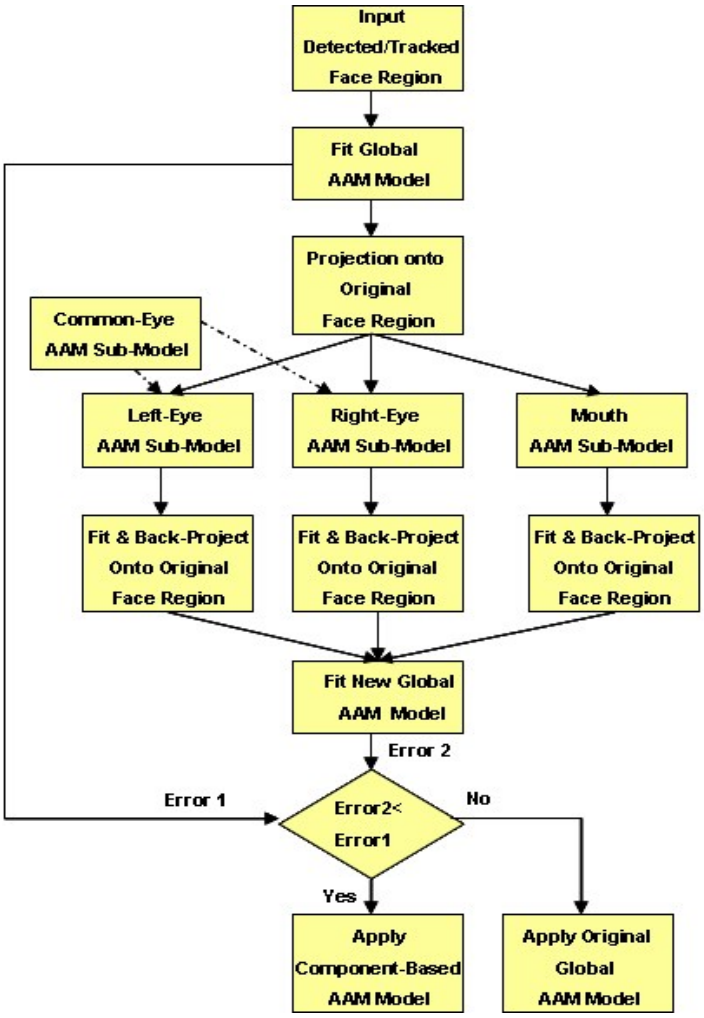


Figure 8.8. Component-based AAM fitting algorithm, aimed to fit the entire facial region.

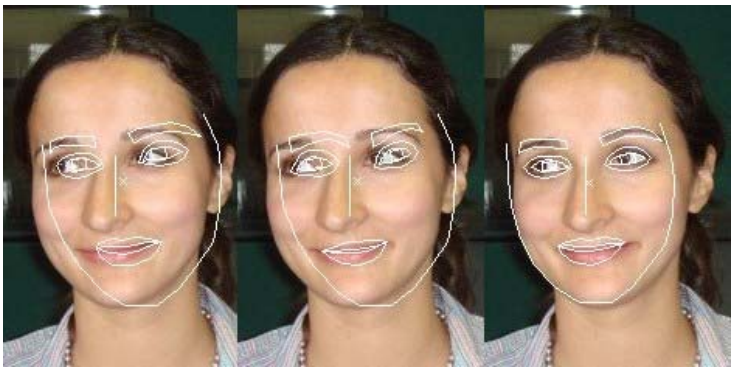


Figure 8.9. Example of shape fitting on an unseen picture with expression and pose. The first is the result of a standard AAM; the second picture represents the fitting of the AAM sub-models, while the third picture depicts the component-based result.

Now, in order to quantitatively evaluate the overall performances of a component-based AAM, we must next measure its accuracy in terms of expression classification/ recognition rates.

Experimental Series 8.11- The standard AAM vs. the component-based AAM in expression analysis

Goal: *compare the performances of a standard AAM global approach, compared to our component-based representation*

Databases: *FEEDTUM, MMI, our own database*

Number of training pictures: *30 for FEEDTUM/MMI, 20 for our database*

Number of tested pictures: *150 for FEEDTUM/MMI, 50 for our database*

Classifier method used: *Euclidean NN, RBF-SVM with* $\delta = 2^2$

AAM model*: a global standard AAM and a component-based AAM for the face region*

Number and type of AAM parameters used for classification: *first 40% of the shape parameters for each trained model*

Tables 8.15-8.17 compare results for expression classification/ recognition rates for a global face model, when compared to a component-based formulation. The first two tables use an NN classifier, while the last is based on SVM classification. In Tables 8-15-8.16, we also restate for comparison the results for the eye and lip features, i.e., the sub-models of our component-based representation obtained in section 8.5.1 and originally presented in Table 8.13.

TABLE 8.15. SUMMARY OF THE SYSTEM ACCURACIES (%) FOR RECOGNISING EXPRESSIONS WITH EUCLIDEAN-NN.

Database	Recognition rate-eyes (%)	Recognition rate-mouth (%)	Recognition rate-face (%)	Recognition component-based (%)
FEEDTUM	27.37	27.37	35.71	54.28
MMI	34.1	26.95	40.88	60
Our database	22.57	27.01	22.66	28.19
Overall	28.01	27.11	33.08	47.49

TABLE 8.16. SUMMARY OF THE SYSTEM ACCURACY (%) FOR CLASSIFYING EXPRESSIONS (NN) USING THE EYE-AREA, THE MOUTH-AREA, THE FACE MODELLED WITH A STANDARD OR WITH A COMPONENT-BASED AAM.

Classifier	FEEDTUM				MMI			
	Eyes	Lips	Face	Comp	Eyes	Lips	Face	Comp
Surprised/ Disg. (%)	44.37	55.87	85.83	83.33	73.12	76.66	85.62	78.33
Surprised/ Happy (%)	28.75	69.37	88.33	83.33	68.75	75	86.25	79.16
Happy/Sad (%)	41.87	46.25	68.33	82.22	72.5	50	73.12	77.5
Surprised/ Sad (%)	33.12	51.66	93.33	83.33	69.37	52.5	77	73.33
Overall (%)	37.02	55.78	83.95	83.05	70.93	63.54	80.49	77.08

Table 8.17 presents the results of the standard AAM (in the columns marked with index 1) and of the component-based AAM (in the columns marked with index 2). The classification rate average when using a component-based AAM is of 73.02%, while for a classical AAM is of 69.43%. The results confirm the improvement brought by a sub-models representation.

TABLE 8.17. SUMMARY OF THE SYSTEM ACCURACIES (%) FOR CLASSIFYING EXPRESSIONS WITH SVM FOR A STANDARD AAM (1) AND FOR A COMPONENT-BASED AAM (2).

	A		D		F		H		N		Sa		Su	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2
A	70	75	63	75	70	75	66.7	68	70	75	70	79.4	73.4	68.7
D			79.4	73.4	63.4	70	70	75	73.4	75	53.4	68	70	71.4
F					72	70	66.7	73.4	56.7	68.7	80	83.4	53.4	63
H							92.7	93.4	73.4	75	60	75	66.7	66.7
N									79.4	68.7	53.4	63	80	83.4
Sa											68.7	63.4	76.7	75
Su													71.4	73.4

8.7. Chapter Summary and Outcomes

In this chapter, we analysed the feasibility of integrating AAM techniques into an automated facial expression analysis system. We provided details of a comprehensive set of experiments on facial expression classification and recognition for still images using the standard AAM formulation. We investigated the effect of reducing the

number of features (model parameters) used for classification, focusing only on expression-specific shape parameters. Using AAM as feature extraction method proved partially successful, even with a simple Euclidean-distance classification scheme. More sophisticated classifiers, such as SVMs, achieved improved results although the overall performance was disappointing.

We then analysed the limitations of this approach and we proposed a component-based AAM facial representation which combined the benefits of a global pose match with localised matching of sub-models to eyes and lips features. This approach has been shown to significantly improve facial expression analysis, as demonstrated by our positive results presented in section 8.6. This methodology for facial expression recognition has been shown to achieve accuracies of up to 83% and we expect that it can be further improved on through a number of additional refinements which are outlined in Chapter 9.

Chapter 9 - Facial Expression Model Refinements

Although the results presented at the end of Chapter 8 are very promising it is clear that additional improvements are desirable in order to achieve a truly robust and reliable facial expression analyser. In earlier discussion we have identified some of the drawbacks which are generic to all AAM models. Consideration of these drawbacks suggests some approaches to further enhance our face model. In this chapter we explore some such approaches, presenting preliminary findings which indicate that further improvements can be achieved on the performance results presented in Chapter 8.

The first topic we consider is the potential to use a person-specific AAM. Such a model eliminates the variability between persons from the statistical model, enabling it to better model other sources of variability. Thus in section 9.1 we present a comparison between person-specific and generic AAMs. While there are advantages to using a model trained explicitly for a particular person, the complexity of this approach does make it more difficult to realise practical solutions as the user must perform training according to quite strict criteria. In this section we shall attempt to quantify the performance differences so that the reader can judge the advantages of a person-specific model.

Then, in section 9.2 we describe how our approaches for still images can be extended to track and recognise expressions in a sequence of images using person-specific AAMs. Applying a model to a sequence of images enables techniques which could track and analyse the transitions between facial expressions, providing an additional level of sophistication. In this section we examine some simple improvements which can be achieved from an analysis of image sequences demonstrating the potential and scope of such dynamic analysis methods.

In section 9.3 alternatives to our basic classifiers are proposed with a view to improve expression analysis rates. Firstly, additional criteria which may be used to classify facial images using AAM features are described in section 9.3.1. These include gender, age, and racial classification. We have performed some preliminary work on using AAM for such criteria and summary details are provided here.

Motivated by the successful preliminary results of gender classification, in section 9.3.2 it is shown how a combination of classifiers can improve facial

expression analysis results. The concept is to use more specific AAM models once a particular criterion has been determined. So once a person's age class is determined we may select an appropriate model for the relevant class: child, youth, young adult, mature adult, senior. Similar logic applies to gender. As in the case of a person-specific model this approach removes one significant source of model variability.

9.1. Expression-Specific vs. Expression-Generic Models

By definition, the expression of a face is the focal element in facial expression analysis. Consequently, identity-related information represents an unwanted source of variability.

There are two modalities of training an AAM, resulting in two types of models [241]: generic AAMs and person-specific AAMs. Depending on the chosen training approach, AAM can contain, or not, information variations which describe aspects of a person's identity. In applications such as facial recognition [82] it is essential to capture, and even seek to emphasise, such identity information. However for the specific purpose of determining facial expression such information is not directly useful, although it will enable the model to adapt to different individuals. This leads us to consider the differences between generic and person-specific AAM models:

G*eneric AAMs*, i.e., models that represent a general variation in appearance, have the advantage that it can model any face, including unseen subjects for the training set. One common application where it is essential to employ such a generic AAM would be face recognition. In such an application the core purpose of the model is to measure the differences between individuals.

Person-specific AAMs model the variations in appearance of a single person across poses, illuminations, expressions and any other sources of variability incorporated into the training set. Such a person-specific AAM might be useful for interactive user interface applications that involve head pose estimations, gaze estimations, or expression recognition. Note that the model really is person-specific and will perform unreliably if it is used to measure the characteristics of another person. The challenge for such models in practical use applications is how to train and verify an accurate model with minimal user input.

In the literature Gross et al. [241] analysed the two models and proved that in the context of the fitting algorithm, the performances of a person-specific AAM are significantly better than the performances of a generic AAM.

In our previous experiments described in Chapter 8, we acknowledged the fact that fitting a single generic AAM on an unseen face under any pose and expression is challenging. The variability of information is really high, so that the fitting process usually gets stuck in a local minimum. In the current section we analyse the influence of the model generality in the context of facial expression recognition.

Experimental Series 9.1- Generic vs. Person-Specific AAM in expression analysis

Goal: *compare generic with person-specific AAMs in the context of expression recognition, so to analyse the influence of the model generality in the context of facial expression variations*

Databases used: *FEEDTUM*

Number of training pictures: *70 (10 for each universal expression and the neutral one)*

Number of tested pictures: *210 (30 for each universal expression and the neutral one)*

Classifier method used: *Euclidean-based NN*

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters used for classification:

1st series- 22 shape parameters for the generic AAM, 4 shape parameters for the person-specific AAM

2nd series- 31 appearance parameters for the generic AAM, 18 appearance parameters for the person-specific AAM

TABLE 9.1. EXPRESSION RECOGNITION RATES (%) OF GENERIC AND PERSON-SPECIFIC AAMS.

Type of model	No. and type of parameters	Recognition rate (%)
Generic	22 shape	40
Person-specific	4 shape	85
Generic	31 appearance	38
Person-specific	18 appearance	94.28

The results of our experiment are summarised in Table 9.1. The person-specific AAM proved to be both *easier to optimise*, i.e., the number of parameters, shape or appearance, is consistently lower, and *more robust to fit* than a generic AAM, by the higher expression recognition rates obtained for a person-specific AAM.

9.2. Expression Classification on Video Sequences- Expression Tracking

Most studies in the literature restrict their analysis of facial expression to still images. As we have shown in this thesis, still images provide sufficient information to recognise many expressions with a reasonable degree of accuracy. They are also preferable from the point of view of the simplicity of their acquisition, manipulation, and analysis.

But we note that facial expressions are dynamic actions. Some expressions, such as agreement and disagreement, require particular details, e.g., head motion, that cannot be determined from a single picture [242]. Also, aspects of the six universal expressions, considered in our previous experiments, can be more easily detected in video sequences than on single images, e.g., eyebrow rising when surprised. Analysing a video sequence of images can provide information about such dynamic features.

Thus if both permanent and transient feature changes could be tracked in an image sequence, the accuracy of analysing facial expression would be improved and the range of expressions which could be determined would be broadened.

To give the reader some initial insights into the potential of dynamic expression analysis from video sequences we present in this section some preliminary experiments on expression tracking.

In our work we simplify the spatio-temporal problem of facial expression variations, by dividing a video sequence into a collection of static snapshots. One may argue that this assumption does not bring any additional information to a static-based system. We consider that by analysing a series of consecutive frames we learn more about:

- Time segmentation, i.e., automatic detection of the time boundaries of expressive sequences.
- Expression intensities, i.e., neutral, onset, offset, and apex states. For each of the universal expressions we can define at least two level of intensity: happiness- ecstasy, anger-rage, fear-terror, sad-grief, disgust-loathing, and surprise-startle.
- Increasing the algorithm speed, as the AAM fitting is applied only on the first frame. For the consecutives ones, shape parameters should only be updated.

The tracking algorithm

In our experiments, for simplification, we used a standard AAM. For the same considerate, only the cases when one face is present in the pictures are taken into account. The first two stages of the tracking algorithm, i.e., initial model fitting and parameters update, were already described in the context of eye tracking (described in section 6.1.3.1) and lip tracking (described in section 6.2.3.1).

Our tracking scheme can be summarised as follows:

In a *first stage*, an *initial model fitting* is done on the first frame of the image sequence. The model is trained to match a specific face to be tracked. It is preferable to use a person-specific model as described in [136]. Firstly, facial parameters, describing shape and texture, are extracted using the AAM approach. This stage is the most computationally expensive and it is performed on the first frame of an image sequence.

Afterwards, in a *second stage*, only the shape parameters are used. These parameters encode head poses and expression variations, as well as position, 2D rotation, and scale parameters. In this stage, the *shape parameters are simply updated*. As the computational requirements are substantially reduced, this operation is now performed on each separate frame.

In a *third stage*, using the updated shape parameters, the *expression variation is tracked and interactively recognised*, using the same classification/ recognition principles as the ones described in section 8.5.

Experimental Series 9.2- Preliminary experiments in expression tracking

Goal: *give a preliminary insight of the performances obtained by a standard person-specific AAM in tracking happiness, i.e., one of the universal facial expressions*

Databases: *FEEDTUM*

Number of training pictures: *person-specific standard AAM trained on 70*

Number of tested pictures: *210*

Classifier method used: *Euclidean-based NN, RBF- based SVM with* $\delta = 2^2$ *trained for happiness*

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters used for classification: *4 shape parameters*

Due to time constraints, we have focused on tracking the happy expression. We tested classification accuracies, time segmentation of the happy state, and its intensity. For the last criteria, we restricted ourselves to just two intensities: happy

and very happy, determined previously by visual inspection. The results are presented in Table 9.2.

TABLE 9.2. SUMMARY OF THE SYSTEM ACCURACIES (%) FOR EXPRESSION TRACKING FOR A PARTICULAR SUBJECT OF FEEDTUM DATABASE.

Classifier	Sequence 1(%)	Sequence 2(%)	Sequence 3(%)
SVM (no intensities)	89.2	82.2	88.2
NN (no intensities)	65.4	65.4	76.8
NN (with intensities)	65.4	65.4	58.3

The results of these initial tests are encouraging. Our tracker based on template updating techniques works well in providing a means for real-time classification of these facial expressions. We remark that while these initial tests are somewhat subjective and limited to a very restricted set of expression classes, our goal here is simply to show the practical potential of expression tracking techniques.

There are issues emphasised by a tracker system. If the fitting on the first frame is not reasonable, caused by limitations of the standard AAM, the errors are transmitted to the following frames, where only shape updates are performed, affecting the classification performances.

Another important observation is that the algorithm works better if the frame rate is higher. Because of the parameters updating stage, the tracker is not able to follow rapid movements or large spatial displacements. If the frame rate is higher, then the motion between each frame is smaller. In consequence fewer iterations are needed per frame, so less time is spent per frame. When processing frames are captured at a low frame rate, the algorithm suffers from being greedy and sometimes it gets stuck in a local minimum. Around 10 frames/second and around two iterations seem to be sufficient for handling normal head motion.

9.3. Improvements of Expression Analysis Rates

In the previous chapters, we analysed ways to improve expression analysis rates, by selecting only relevant facial features/parameters (sections 8.2-8.3), proposing improvements of the facial representation (section 8.6.2), or provide best settings for the classifiers (section 8.5).

In the current section we analyse the impact of a combination of classifiers aiming to recognise facial expressions. We start by investigating other classification

criteria using AAM features, in section 9.3.1. Then, in section 9.3.2, we use these criteria together with our expression classifiers to accurately classify facial expressions.

9.3.1. Other Classification Criteria which can use AAM Features

Although our main research presented in this chapter is focussed on classification of human expressions using AAM we note that the same techniques can be employed to make other determinations from a person's face. As examples we mention age [160, 243], gender [107, 160], and race classifications [244], each of which have been actively researched in the literature.

Note that to the best of our knowledge, there is no prior work addressing the problem of race classification using AAM features. To show the flexibility of our AAM based approach we now present some illustrative tests for each of these criteria.

Experiment 9.3- Series 1: Race classification using AAM features

Goal: *test the AAM discriminative power to classify between races, i.e., Asian vs. Non-Asian*
Databases used: *FERET for Asian faces, FEEDTUM for Non-Asian faces*
Number of training pictures: *30*
Number of tested pictures: *150 (not used in the training set)*
Classifier method used: *Euclidean-based NN, RBF- based SVMs with* $\delta = 2^2$
AAM model*: global standard AAM for the face region*
Number and type of AAM parameters: *18 shape, 22 appearance parameters*

Experiment 9.3- Series 2: Age classification using AAM features

Goal: *test the AAM discriminative power to classify between ages, i.e., child vs. adult*
Databases used: *personal pictures for child faces, FEEDTUM for adult faces*
Number of training pictures: *30*
Number of tested pictures: *150 (not used in the training set)*
Classifier method used: *Euclidean-based NN, RBF- based SVMs with* $\delta = 2^2$
AAM model*: global standard AAM for the face region*
Number and type of AAM parameters: *17 shape, 20 appearance parameters*

Experiment 9.3- Series 3: Gender classification using AAM features

Goal: *test the AAM discriminative power to classify between gender, i.e., male vs. female*
Databases used: *FEEDTUM*
Number of training pictures: *30*
Number of tested pictures: *150 (not used in the training set)*

Classifier method used: *Euclidean-based NN, RBF- based SVMs with* $\delta = 2^2$

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters: *18 shape, 19 appearance parameters*

The experiments in this section are simplified, aiming just an overview of the capabilities and the broadness of the AAM information. The results, summarised in Figure 9.1, underline the superiority of SVM and show that the appearance parameters are indispensable in terms of gender, age, and race information. Now while for expression, shape carries the relevant information, for gender, age, and race description, ***appearance parameters***, i.e., merged shape and texture, proves to be more valuable.

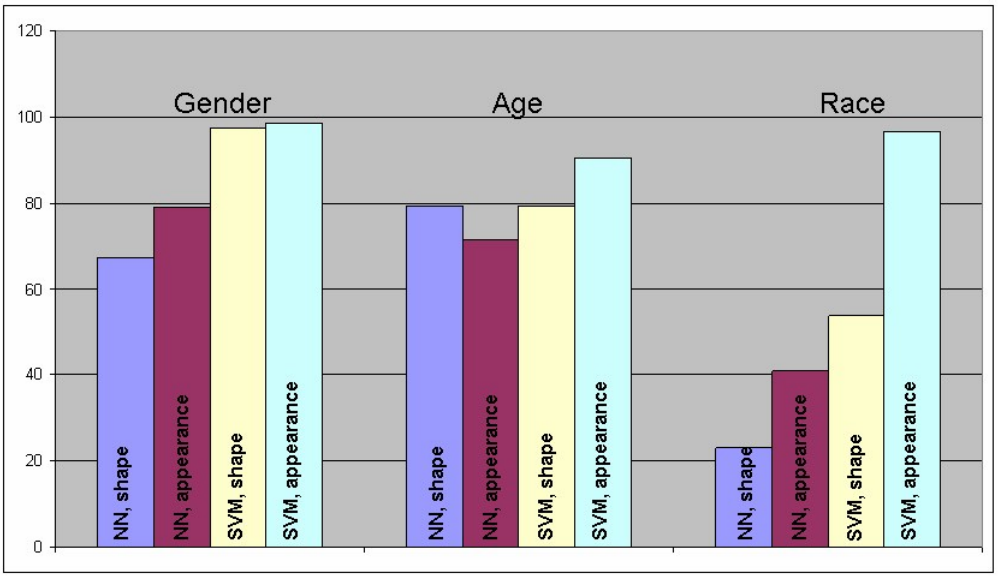


Figure 9.1. Accuracies (%) for Gender, Age, and Race classifiers.

9.3.2. Combination of Classifiers

From the results of our expression classification/ recognition experiments described in section 8.5, it is clear that for our application SVM significantly outperforms NN-based classifiers. Now, a combination of classifiers may be able to further improve these initial classification results [107, 245].

In literature [107, 245] it was experimented that it is easier to find several relatively good classifiers, than finding one single highly discriminant classifier. Another advantage of combining classifiers is that they are less probable to over fit, by reducing the variance of the decision function [246] and they require less computational time [245]. Also, the deficiencies of one classifier can be compensated

by the other classifiers [245], e.g., in our case, one classifier may only poorly distinguish between anger/sadness, but if one of the other classifiers provides very strong discrimination between these two expressions, then the combination of these two classifiers will help resolve the deficiency of the first classifier.

There are many ways to combine classifiers and there is nearly no limit in inventing other combination rules. Meynet researched in his thesis [245] the effect of combining classifiers and modalities to combine them. Between the approaches that the author of [245] proposes, we mention:

- Combine different learning algorithms instead of only keeping the one giving the lowest estimation of the expected risk.
- For a given learning algorithm, combine several candidate solutions. Several candidates can be obtained by various initialisation procedures or various model selection strategies.
- Combination of several classifiers can be obtained from different sources (different training patterns or different features).
- Different training sets can be collected at different times or in different conditions.

In our experiments, we limit ourselves to combining two different classifiers designed for distinct tasks. Motivated by the good results obtained by our gender classifier (section 9.3.1), we test the benefits brought by using a gender classifier previous to performing expression classification, i.e., test "male" and "female"-specific sets of expression classifiers. This idea was already proposed and successfully used in [6].

Features extracted by a trained AAM are used to construct SVM classifiers for four expressions, i.e., happy, angry, sad, neutral. These classifiers are arranged into a cascade structure in order to optimise overall recognition performance. We note that in [160], it is stated that there are gender-specific differences in the appearance of facial expressions that can be exploited for automated recognition. The results we now present corroborate these findings and that cascades are an efficient and effective way of performing multi-class recognition.

Experiment 9.4- Combination of classifiers

Goal: *experiment combinations of classifiers aiming to increase the expression recognition rates*

Databases: *FEEDTUM, MMI*

Number of training pictures: *30 for each classifier*

Number of tested pictures: *150 for each classifier*

Classifier method used: *cascaded RBF-SVM classifiers with* $\delta = 2^2$

AAM model*: global standard AAM for the face region*

Number and type of AAM parameters: *6 shape parameters for FEEDTUM, 7 shape parameters for MMI*

A gender SVM is applied prior to applying a SVM cascade of expression classifiers, resulting in a gender-specific expression recognition scheme. The cascade workflow and the obtained results are summarised in Figure 9.2. We remind the reader that in section 8.5.2.2 a similar SVM cascaded structure was proposed, without the gender/race related discrimination stage. 59.8% expression recognition rate was obtained on the MMI and 64.8% on the FEEDTUM database, as compared to 63.7% and respectively 68.9%, when gender-related information is used.

So, we conclude that a set of pre-filters, i.e., a gender discriminator pre-filter in our experiment, lead to more specifically trained cascades; moreover this has the effect of pushing recognition accuracy levels over 90%. Based on the positive results obtained in section 8.6, we believe that the rates can be improved even more using our proposed component-based AAM representation for facial expression.

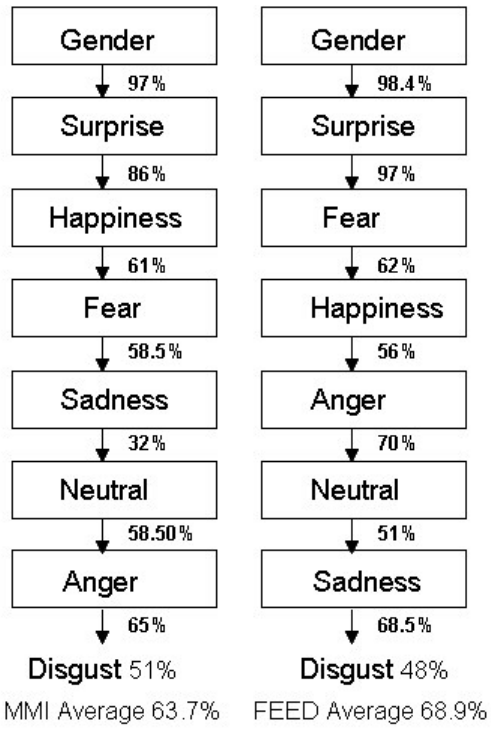


Figure 9.2. Performances of expression recognition of SVM classifiers in a cascade structure, including gender classifications for the MMI and FEEDTUM databases.

Chapter 10 - Conclusions and Future Research

This chapter provides an opportunity to stand back and review the research problems presented, the solutions proposed and implemented, and the experimental results presented in this thesis.

In section 10.1, we summarise our work, emphasising the outcomes and conclusions which resulted.

The main contributions of our work are listed in section 10.2, while in section 10.3, based on the research undertaken for this thesis we identify new problems and propose research directions for those who wish to build on our contributions.

We end this chapter by listing in section 10.4 the most relevant publications that resulted from the work described in this thesis.

10.1. Research Summary

The research undertaken in this thesis represents strategic medium-term research for the sponsoring company, FotoNation (Ireland) Ltd. This was originally focussed on face analysis and recognition techniques, and in particular on AAM face models which represent practical technology that both my supervisor and the sponsoring company considered as having potential for commercialisation as a second-stage development of its in-camera face tracking technology. As the research progressed it was realised that more sophisticated applications than basic face recognition could be realised using AAM modelling techniques.

Thus the core research theme of this doctoral thesis began to emerge. After our initial successes with improving the underlying AAM face model it was decided to pursue studies of some more ambitious and sophisticated models to determine the practicalities of modelling of unconstrained faces with sufficient accuracy to enable a determination to be made of facial expressions.

Thus the eventual end goal of our research became the design and implementation of a working facial expression recogniser based on enhancements of standard AAM techniques.

10.1.1. Review of our Work

The starting point of our research was to survey the literature in order to find an optimal face modelling technique. This was defined as one capable of modelling unseen faces and facial expressions. Following our literature study, we concluded that AAM based techniques presented the best trade-off between efficiency and performance for both face and facial expression modelling. They also offered a practical scaling to embedded devices.

Tests on the standard AAM model used in unconstrained conditions and applied to faces with unseen facial expressions were, however, less than satisfactory. From this work we determined several limitations of the standard AAM technique. After further study and review of the literature we proposed several solutions which should compensate for these limitations.

Firstly, we verified which facial features contribute the most in depicting facial expression. We found that the eyes and lips are the most significant. Thus, we adapted the standard AAM to model eyes and then lips, as detailed in Chapter 5. We tested optimal shape annotations, training sets, or allowed variances for shape and texture in each case. We next developed shape patterns that accommodate these features in all their states, e.g., (i) open and closed eyes in the same eye template, (ii) neutral or smiling lips in the same lips template. We concluded that the standard AAM formulation, although successful in the case of neutral faces, still exhibited limitations when modelling individual facial features and facial expressions.

Specific challenges for the eye model included unseen head poses and expressions, occlusions of one or more components of the eye model, or differences in expression between the two eyes, e.g., winking. To address these challenges we further developed our eye model into a component-based AAM eye model. We adapted this approach for the eye region, by combining a two-eye global model with a single-eye sub-model, as described in Chapter 6. This approach benefited from both the generality of the global AAM and the local optimisations provided by its sub-models. Then, we proposed, implemented and verified several practical applications for our novel eye model, i.e., a blink detector and an eye tracker.

With regard to the lips feature, we concluded that the main cause for failures of the standard AAM is the poor model initialisation. This was mainly due to the weak lip colour contrast and significant overlap in colour features with the face

region. So, we improved the standard AAM formulation by applying a pre-processing step, i.e., a hue filter that offers more accurate initialisation of the lip region, as described in Chapter 6. In addition we proposed a practical application of the novel lip model - a smile detector.

We next undertook detailed testing and comparison of our modelling approaches. We analysed their performance by comparing them with standard AAMs or by testing them in the context of several consumer applications. We concluded that for all tasks, better results were obtained in the case of our model enhancements.

The model enhancements above were then studied in the context of an automatic face expression recognition system, as described in Chapters 8 and 9. We reported results on a series of experiments comparing feature selection methods and recognition engines. Using AAM as feature extraction method was acceptable, even with a simple Euclidean-distance classification scheme. More sophisticated classifiers, such as SVMs, performed significantly better. Overall, this initial work has shown the potential of AAM as a means for determining robust feature sets for expression classification.

We investigated the effect of using a reduced number of features for classification, focusing only on expression specific shape parameters. Irrelevant features in the vector increased classifier complexity, decreasing performance in the majority of cases.

We also determined that the accuracy of feature alignment is a key determining factor for system performances. The goal of face alignment is firstly to accurately locate facial features such as eyes, nose, mouth, and face outline, and secondly to normalise shape and texture accordingly. Motivated by the amount of information contained in these facial features and by the good results obtained for expression decoding we developed a detailed face representation based on facial sub-models.

The proposed approach towards facial expression recognition has been evaluated to achieve accuracies of up to 83%. The adopted technique is tailored to cater to both genders and a variety of races. We further enhanced and proved the idea that performances can be further improved by combining gender and expression classifications. In our work, we also addressed the problem of dynamic classification. Finally, a preliminary evaluation of the use of video sequences showed that additional discriminative power could be added to the classification task.

10.2. Principle Outcomes of this Research

10.2.1. Main Contribution

There are a number of principle contributions arising from this research:

Development of an eye model using variations of AAM robust to blink (described in Chapters 5, 6 and published in [233, 235])
We have adapted AAM techniques to provide a model of the eye region which can describe open, closed, and intermediate eye-states. The eye model employs the concept of overlapping key-points which enables our model to handle both open, close, and intermediate eye-states. A patent application (currently unpublished) directed to this aspect of the model has been filed.

Although AAM was used before to model the eye appearance and gaze [3, 4], the authors of [208, 211] could not successfully model eye-blink. We propose a solution that models eye blink, and can be readily adapted to determine eye-gaze and wink.

We later refined the technique by applying a component-based AAM, previously used to model faces [94], adapted now to details of the eye area.

An improved component-based AAM model adapted for facial expressions (described in Chapter 8 and proposed to be published)
A component facial AAM model is introduced employing both a global model and local sub-models based on the eye and lips models of Chapter 6. Although component AAM models are known in the literature [94] their use in facial expression analysis with our specialised eye and lips models represents an improvement over known techniques.

10.2.2. List of Other Contributions

Other additional contributions resulted from our work can be named:

Development of a lip model using AAM and pre-processing techniques (described in Chapters 5, 6 and published in [234])
It consists of the development of a lip model using a combination of AAM and hue-filtering pre-processing techniques.

Previously to our research, AAM was used by other researchers [232] to model the lips area. However, we propose an improved approach, by using a hue filter as a pre-processing technique for a more robust initialisation of the model together with AAM. The combination of the two techniques is a novel approach that proved able to increase the AAM fitting rates (section 7.2).

Development and implementation of a series of applications based on our AAM sub-models (described in Chapter 6 and published in [233-235])
Novel approaches for blink/ smile detections and for eye/lip tracking are proposed, based on our improved facial sub-models. These techniques were not previously used to develop such applications.

Choice of optimal AAM features/ parameters for expression analysis (described in Chapter 8 and proposed to be published)
The most relevant AAM features/ parameters are analysed in the context of expression analysis. We investigated which type of AAM feature, i.e., shape, texture, appearance, contributes the most in modelling facial deformations caused by facial expressions. Although similar studies exist in literature [68, 97, 140], opinions are divided among researchers; consequently, we searched for the relevant feature for our framework.

Then, we investigated the optimal AAM parameters for modelling facial expressions and we discard the irrelevant or redundant ones. Our proposed design used the subset of model parameters which mainly encode expression information, discarding thus the information related to variability in pose and identity, leading to more accurate expression recognition rates. Similar approach was already proposed by us in [8], adapted though for face recognition use.

Evaluation of the expression discrimination power of independent facial features, such as the eyes and the lips (described in Chapter 8)
Although the results do not motivate us to propose an expression analysis system based only on the eyes or lips facial features, this study brings novel information concerning the individual usefulness of these features in an automated expression classification system and motivates our component AAM facial expression approach.

Elaboration of a reduced database containing concomitantly variations in pose and in facial expressions, including blink and gaze, useful for our tests (described in Appendix B)

As we noticed that the literature lacks in databases presenting a range of facial deformations under several head poses and illuminations, we proposed our own collection of pictures, incorporating concomitantly both pose and the six universal expressions and the neutral one, or blink and gaze actions. The database is described in Appendix B.

10.3. Future Directions

The work presented in this dissertation showed areas for interesting future research:

An adaptive hue filter

In Chapter 6, we have used a hue filter as a pre-filtering method in order to initialise the lips model. The filter parameters were calculated based on statistical characteristics of the constrained conditions of a set of images. Our work could be further improved through the implementation of an adaptive hue filter capable of changing its parameters to match the original acquisition conditions of each image.

Robustness to illumination variations

In Appendix D, we conducted preliminary experiments in which we incorporated an AAM algorithm extension aiming to increase robustness to directional illumination variations. We demonstrated that applying such an algorithm at the training stage increases the fitting accuracies of the resulting AAM model. The increase in convergence accuracy has a direct, positive impact on the expression recognition rates of our system.

In the future, more robust and efficient solutions should be incorporated in order to solve robustness to unconstrained illumination conditions.

Robustness to partial facial occlusions

The problem of occlusion, although a classical limitation of AAM, was not directly studied in this thesis. A robust solution directly approaching this issue should be implemented into our system to be able to decode partially occluded facial

expressions. One approach may be through using sequences of images rather than a single still image; temporary spatial occlusions could then be handled in a number of ways by tracking the face (or other object) across multiple image frames and by extrapolating the face in an occluded facial from earlier or later non-occluded frames.

Towards real-time applications

Our system is not developed with the explicit purpose of performing in real-time, but we have studied some optimisations which could significantly improve performance in real-time embodiments. One such optimisation is a reduced number of input features. We have shown that as few as 6 shape parameters can provide the highest model performance while providing at least a 4x speed up.

Extending expression recognition to the FACS system

Our system should be extended to take into account other categories and combinations of emotions, like the FACS system (defined in section 2.3.1).

10.4 Summary of Relevant Publications

After the completion of each stage in our work, we published papers to report our research progress to the academic world. A list of published or submitted academic material is presented in the following:

Conference papers & associated presentations

- M.C. Ionita, I. Bacivarov, P. Corcoran, Separating Directional Lighting Variability in Statistical Face Modelling based on Texture Space Decomposition, DSP, Cardiff, Wales, June 2007.

The paper proposes an AAM-based face representation in which two sets of parameters are used to control inter-individual variation and separately intra-individual variation due to changes in illumination conditions. The idea is applied and tested in Appendix D of this thesis in the context of facial expression analysis. The main author is responsible for the main idea of the paper, while I participated with the survey of the literature, wrote Section 1 of the paper, executed some of the experiments, and gave valuable comments.

- M.C. Ionita, P. Corcoran, I. Bacivarov, Next Generation Face Tracking Technology Using AAM Techniques, ISSCS, Iasi, Romania, July 2007.

The paper proposes solutions to problems related to the implementation of AAM in real-time face tracking applications for embedded systems. Another contribution is a new AAM training and model construction approach with increased robustness to head pose and illumination variations. The training approach and tracking schematic are described and used for our eye and lip models (Chapter 6) and for our face model dedicated to expression analysis (Chapters 8, 9). The main author is responsible for the main idea of the paper. I participated with the survey of the literature, wrote Section 1 of the paper, executed some of the experiments, and gave valuable comments.

- P. Corcoran, I. Bacivarov, M.C. Ionita, A Statistical Modeling Based System for Blink Detection in Digital Cameras, ICCE, Las Vegas, January 2008.

The paper describes initial experiments of an AAM eye model robust to blink (see Chapter 6 of this thesis). The idea is extended in the corresponding journal paper (no.6 in this list). From our knowledge, AAM was not used to model blink up to date. The first author is responsible for the idea of using AAM to model eyes and other valuable comments. I am responsible for the novel approach, the experiments, and the paper writing. I also received feedback from the third author, who also helped with some aspects of the model practical implementation.

- I. Bacivarov, M.C. Ionita, P. Corcoran, A Combined Approach to Feature Extraction for Mouth Characterisation and Tracking, ISSC, Galway, June 2008.

AAM and hue filter pre-processing techniques are coupled and adapted to model and track lips. From our knowledge, AAM was not used to model lips up to date. Complete description and testing of the approach are described in Chapters 6-7 of this thesis. I am responsible for the idea, the experiments, and the writing of this paper. The other authors participated with valuable feed-back.

- M.C. Ionita, P. Corcoran, I. Bacivarov, Smart Cameras: 2D Affine Models for Robust Person Recognition in Consumer Images, ICCE, Las Vegas, January 2009.

An initial proof-of-concept using AAM to perform face recognition which is robust to pose and illumination variations is presented. The ideas presented in this paper are adapted also in the context of facial expression recognition in Chapters 8-9. My contribution to this paper is related to the execution of some experiments, and other valuable comments

Journal papers

- I. Bacivarov, M.C Ionita, P. Corcoran, Statistical Models of Appearance for Eye Tracking and Eye-Blink Detection and Measurement, in the IEEE Transactions of Consumer Electronics, 2008.

AAM is developed to track and detect eye blink, robust to variations in head pose or gaze. This global model is further extended using a series of sub-models to enable independent modelling and tracking of the two eye regions. Several methods to enable measurement and detection of eye-blink are proposed and evaluated. Complete description and testing of the approach are described in Chapters 6-7 of this thesis. I am responsible for the idea, the experiments, and the writing of this paper. The others authors participated with valuable feed-back.

Submitted papers

- I. Bacivarov, P. Corcoran, M. C. Ionita, Smart Cameras: 2D Affine Models for Determining Subject Facial Expression, in the IEEE Transactions of Consumer Electronics, 2009.

An initial proof-of-concept using a component-based AAM formulation to measure facial parameters and a set of SVM classifiers to determine facial expressions are described. The schematic is designed as to be robust to variations in pose and illumination and able to perform in real-time. Complete description and testing of the approach are described in Chapters 8-9. I am responsible for the idea, the experiments, and the writing of this paper. The other authors participated with valuable feed-back.

Patent disclosures and filings

- 20090003661- M.C. Ionita, I. Bacivarov, P. Corcoran, Separating a Directional Lighting Variability In Statistical Face Modelling Based On Texture Space Decomposition 01-01-2009.
- 20080205712- M.C. Ionita, I. Bacivarov, P. Corcoran, Separating Directional Lighting Variability in Statistical Face Modelling Based on Texture Space Decomposition 08-28-2008.

These disclosures are based on paper number 1.

Papers in preparation

- P. Corcoran, I. Bacivarov, Facial Expression Modelling for Gaming Applications, to be submitted for the ICE-GIC 2009, and an extended version for the IEEE CE Transactions.

The paper proposes solutions to problems related to the implementation of facial expression AAM solutions in real-time gaming applications for embedded systems. The paper is based on experiments developed in Chapters 8-9 of this thesis. The main author participated with the idea of using our model in gaming applications and partially wrote the paper. I am responsible for the ideas, experiments, and I wrote parts of the paper.

- I. Bacivarov, P. Corcoran, Person-Specific vs. Generic AAMs in facial expression analysis and tracking, to be submitted for the Pattern Recognition Letters.

The paper proposes a comparison of the two approaches in the context of facial expression analysis and tracking, from the point of view of building and fitting the models and facial expression recognition performances. The paper is based on experiments developed in Chapters 8-9 of this thesis. I am responsible for the idea, the experiments, and the writing of this paper. The second author participated with valuable feed-back.

Appendix A - Principal Component Analysis

Our short description of the Principal Components Analysis (PCA) algorithm is based on the complete tutorial [247] of Smith. PCA is an unsupervised subspace extraction method. The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. This is achieved by transforming to a new set of variables the principal components which are uncorrelated and which are ordered so that the few first retain most of variation present in all original variables. The algorithm is formed of several steps:

Step 1: Data collection

For simplicity, in our current description we presume that the collected data have only two dimensions, *x* and *y*.

Step 2: Subtract the mean

For PCA to work properly, we have to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the *x* values have $\bar{x}$, i.e., the mean of the *x* values of all the data points, subtracted, and all the y values have $\bar{y}$ subtracted from them. This produces a data set whose mean is zero.

Step 3: Calculate the covariance matrix

Then, the data covariance has to be calculated. Covariance is always measured between two dimensions, using the following formula:

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \qquad \text{(A.1)}$$

where *n* refers to the number of elements in the data set.

If we have a data set with more than two dimensions, there is more than one covariance measurement that can be calculated. In fact, for a *n*-dimensional data set, we can calculate *n!/(n-2)!*2* different covariance values.

A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix. So, the definition for the covariance matrix for a set of data with n dimensions is $C^{n*n}=(c_{ij})$, where $c_{ij}=cov(Dim_i, Dim_j)$, C^{n*n} is a matrix with n rows and n columns, and Dim_i is the i^{th} dimension.

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

Since the covariance matrix is square, we can calculate the eigenvectors and eigenvalues for this matrix. These are rather important, as they tell us useful information about our data. By the process of taking the eigenvectors of the covariance matrix, we have been able to extract lines that characterise the data. The rest of the steps involve transforming the data so that it is expressed in terms of them lines.

Step 5: Choosing components and forming a feature vector

In this step we produce data compression and dimensionality reduction. In general, if we look at the eigenvectors and eigenvalues of a set of data, we notice that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the highest eigenvalue is the principle component of the data set.

In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives us the components in order of significance. Now, we can decide to ignore the components of lesser significance. To be precise, if we originally have n dimensions in the data, we calculate n eigenvectors and eigenvalues, and then we choose only the first p eigenvectors, then the final data set has only p dimensions.

Step 6: Deriving the new data set

This is the final step in PCA, and is also the easiest. Once we have chosen the components, i.e., eigenvectors, that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

$$FinalData=RowFeatureVector*RowDataAdjust \qquad (A.2)$$

where *RowFeatureVector* is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top. *RowDataAdjust* is the mean-adjusted data transposed, i.e., the data items are in each column, with each row holding a separate dimension. *FinalData* is the final data set, with data items in columns, and dimensions along rows

PCA has many applications in Computer Vision, such as finding patterns, or image compression.

Appendix B - Description of Databases used in our Work

Our own database of Expression& pose, blink, and smile

For some of our experiments, we collected a series of pictures and video sequences depicting facial expressions and pose, or blink, gaze, and smile. Our collection emulates conditions of consumer pictures, as they embrace different lighting conditions, subjects, or head poses. Examples of our collection of pictures are given in Figure B.1-B.2.

FERET

The colour FERET database [165, 166] was collected between December 1993 and August 1996. In 2003 DARPA released a high-resolution, 24-bit colour version of these images. The dataset tested includes 2413 still facial images, representing 856 individuals. Some sample images from the colour version of the database are presented in Figure B.3.

MMI

The MMI database [167] contains still images and videos displaying both the six basic emotional displays and the individual activations of each of 44 existing facial muscles. The database holds over 2000 videos and over 500 images of about 50 subjects displaying various facial expressions on command. Along with the data samples, metadata in terms of displayed AUs is available. The database represents a number of demographic variables including ethnic background, gender, and age. Although it contains a large number of variables, i.e., six emotions, combinations of 44 AUs, various human aspects, the database is easily searchable. Examples of the MMI database are presented in Figure B.4.

FEEDTUM

The FEEDTUM Database [169] with Facial Expressions and Emotions from the Technical University of Munich is an image database containing face images showing a number of subjects performing the six different basic emotions defined by Eckman and Friesen. The database has been developed in an attempt to assist researchers who

investigate the effects of different facial expressions. The database contains material gathered from 18 different individuals. It is intended to expand this gallery in the future. Each individual performed all six desired actions three times. Additionally three sequences doing no expressions at all are recorded. Altogether this gives an amount of 399 sequences. Depending on the kind of emotion, a single recorded sequence can take up to several seconds.

An important aspect of the database is that consists of elicited spontaneous emotions. To elicit the emotions as natural as possible, the authors decided to play several carefully selected stimuli videos and record the participants' reactions. For this purpose a video monitor together with a mounted camera on top were employed, which enables a direct frontal view. Both devices were controlled by a dedicated software that induced the desired emotions and started the recordings at the expected times. Samples from the FEEDTUM database are presented in Figure B.5.

Georgia Tech Face Database

Georgia Tech face database [238] contains images of 50 people taken in two or three sessions between 06/01/99 and 11/15/99 at the Centre for Signal and Image Processing at Georgia Institute of Technology. All people in the database are represented by 15 colour JPEG images with cluttered background taken at resolution 640*480 pixels. The average size of the faces in these images is 150*150 pixels. The pictures show frontal and/or tilted faces with different facial expressions, lighting conditions, and scale. Each image is manually labelled to determine the position of the face in the image. Samples from the Georgia Tech Face database are presented in Figure B.6.

VidTIMIT Video Dataset

The VidTIMIT database [239] is comprised of video and corresponding audio recordings of 43 people, reciting short sentences. The sequence consists of the person moving their head to the left, right, back to the centre, up, then down and finally return to centre. The images are stored as JPEG files with a resolution of 512*384 pixels. Samples from the VidTIMIT database are presented in Figure B.7.

Figure B.1. Samples from our collection of pictures concomitantly depicting expression and pose.

Figure B.2. Samples from our collection of pictures concomitantly depicting blink, gaze, wink, in various illuminations and head poses.

Figure B.3. Samples from the FERET database.

Figure B.4. Samples from the MMI database; each column depicts a specific facial expression: from left to right happiness, fear, anger, surprise, sadness, disgust, and neutral face.

Figure B.5. Samples from the FEEDTUM database: each row presents a specific subject depicting the seven universal facial emotions.

Figure B.6. Samples from the Georgia Tech Face Database.

Figure B.7. Samples from the VidTIMIT Video Dataset.

Appendix C -Applications of our Models-The "Eyes-Wide-Open" Application

In section 6.1.3.2 we have described the principle of a blink detector. A blink detector would be very useful in digital cameras. If a camera has a blink detector, this device would alarm the camera if a subject has his/ her eyes closed. If a blink flag is raised, the camera could either wait until all subjects have their eyes open or reshoot. So a blink detector would ensure that the camera grabs a shot when all subjects are wide-eyed, embellishing our collection of personal pictures. The benefits of such a camera would be particularly significant for parents with several children who can never get that perfect group shot!

Other potential applications of such a model would enable post-acquisition manipulation of an image, or a sequence of images, super-imposing an adjusted eye region over the original image. This would enable undesirable blinking events to be corrected in an image, or conversely an eye-wink could be inserted into an image to enhance its impact or attractiveness.

If the photo is already taken and one has closed eyes, we still want to be able to correct it. It would be useful to have an application, embedded or not in the digital camera, that automatically open one's eyes in the picture. Our eye-model (described in section 6.1) enables a replacement to be made of the subject's closed eyes with equivalent open-eyes. Two pictures are used as input: the blinked photo to be fixed and a picture of the same subject, having its eyes open. The pictures including open eyes can be obtained from prior information in the camera, for instance by reviewing already existing photos.

The principle is graphically exposed in Figure C.1. Our methodology is applied as follows: an AAM fitting is performed on both pictures, i.e. with open and closed eyes. The procedure enables segmenting the eyes in the two photos and retrieving their shape, pose, and texture parameters. The original texture for the open eyes is transferred into the closed eye photos, respecting the closed eye pose parameters, i.e. translation, rotation, and scaling. For more realistic results, the eyebrow textures are discarded. Some initial experiments are performed. The main requirement for our application is that both the open-eye and closed-eye shapes are fitted highly accurate. The results can be judged from Figure C.2.

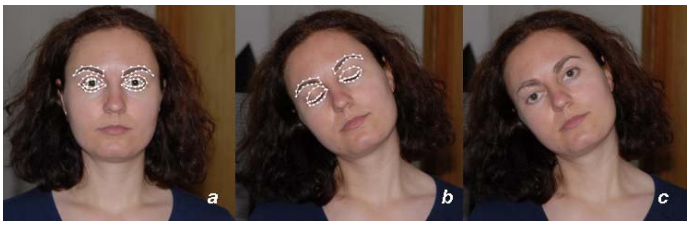


Figure C.1. The original pictures with a. open, b. closed eyes, and c. the processed picture, where closed eyes were replaced by open eyes.

Figure C.2. Examples of the original pictures with closed eyes-upper row and the resulting pictures after the "eye-opening" process-lower row. The examples were chosen from the pictures for which AAM found the eye parameters correctly.

Appendix D - Robustness to Illumination Challenges

Illumination robustness- The illumination directional model

In order to deal with ***illumination*** problems, we use an earlier technique that we proposed in [83]. For a more detailed description of the algorithm, please refer to the thesis of a fellow researcher, M.C. Ionita [82].

It is known that PCA, used by AAM to model shape and texture, cannot differentiate between various sources of texture variability. A special technique to process the illumination is used along the standard method to build a statistical model. The texture space is separated into two subspaces, one for inter-individual variability and another for variations caused by directional changes of lighting conditions. The model is built in the following manner:

- A shape model and a texture model are built from different individuals at constant frontal illumination.

- A set of images with various directional lighting conditions are filtered by projecting the corresponding textures on the previously built space; the residues are used to build a second subspace for directional lighting.

The two texture subspaces are then reintegrated into a single texture model. The resulted texture space allows a more advanced control on the two sources of facial variability represented by identity and directional lighting. The resulting subspaces are orthogonal, so the overall texture model can be obtained by a simple concatenation of the two subspaces.

The main advantage of this representation is that two sets of parameters are used to control separately inter-individual variation and intra-individual variation due to changes in illumination conditions. It significantly improves the fitting stage, the improvements being transmitted to the expression decoding stage, too.

Some preliminary tests have been performed in order to evaluate this approach in the context of expression decoding. Although this aspect is not the main focus of this research, and exact figures are not provided, it is believed, and also supported by our preliminary tests, that this algorithm can add accuracy in cases of directional illumination.

Texture Normalisation

Texture normalisation can help to improve the performances of an AAM fitting and, subsequently, the accuracy of decoding expression information [82]. This research demonstrated that the use of an on-channel-normalisation instead of a global normalisation significantly improves fitting accuracy and AAM convergence when dealing with colour images.

In [82] it was shown that the RGB colour space has high inter-channel correlations and that an improved colour AAM can be obtained if each image channel is decorrelated from the other image channels and individually normalised. This can be realised by employing the I1I2I3 colour space, originally proposed by Ohta et al. [248].

Experimental Series D1- On-channel texture normalisation

Goal: *test the texture normalisation effects, standard vs. on-channel, from the point of view of expression recognition rates*

Databases: *FEEDTUM*

Number of training pictures: *70 for MMI, 70 for FEEDTUM, 35 for our own database*

Number of tested pictures: *210 for MMI, 210 for FEEDTUM, 70 for our own database*

Classifier method used: *Euclidean-based NN*

AAM model*: component-based AAM for the face region*

Number and type of AAM parameters used for classification: *18 shape parameters for FEEDTUM, 17 shape parameters for MMI, 17 shape parameters for our own database*

Here we present some basic experiments performed to confirm this theory and its applicability to facial expression modelling. The figures refer to expression recognition when applying a Euclidean-based NN classifier. The results are slightly better when applying on-channel normalisation (31.9%) than the standard global normalisation approach (30.4%). This type of normalisation is useful especially when dealing with unconstrained scenarios that contain variations in illumination.

TABLE D.1. A COMPARISON OF EXPRESSION RECOGNITION ACCURACY WHEN USING A STANDARD TEXTURE NORMALISATION APPROACH AND AN ON-CHANNEL TEXTURE NORMALISATION APPROACH.

Normalisation	MMI (%)	FEEDTUM (%)	Our database (%)	Average (%)
Standard	32.14	38.09	20.96	30.4
On-channel	33.85	38.09	23.70	31.9

References:

1. http://www.embedded-computing.com/news/db/?10084.
2. Corcoran, P. and G. Costache, *Automated sorting of consumer image collections using face and peripheral region image classifiers.* IEEE Transactions on Consumer Electronics, 2005. **51**(3): p. 747-754.
3. Yang, M., et al., *Face detection for automatic exposure control in handheld camera.* Computer Vision Systems, 2006 ICVS '06, 2006: p. 17 - 25.
4. Casselberry, J., *Facial Expressions of Pain in Elderly Adults with Dementia.* Journal of Undergraduate Research, March/April 2006. **7**(4).
5. *Assessing and Treating Pain in Older Adults.* American Medical Association, 2007.
6. Ishikawa, T., et al., *Passive Driver Gaze Tracking with Active Appearance Models.* Technical Report, 2004.
7. Ekman, P., W. Friesen, and M. O'Sullivan, *Smiles when lying.* Journal of Personality and Social Psychology of Women Quarterly, 1988. **54**: p. 414-420.
8. Ionita, M., P. Corcoran, and I. Bacivarov. *Smart Cameras: 2D Affine Models for Robust Person Recognition in Consumer Images*. in *International Conference on Consumer Electronics*. 2009. Las Vegas, USA.
9. Belhumeur, P.N., J. Hespanha, and D.J. Kriegman, *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. **19**(7): p. 711-720.
10. Blanz, V. and T. Vetter. *A morphable model for the synthesis of 3D faces*. in *Computer Graphics Annual Conference Series (SICGRAPH)*. 1999.
11. Cootes, T.F., K.N. Walker, and C.J.T. . *View-based active appearance models*. in *4th International Conference on Automatic Face and Gesture Recognition*. 2000. Grenoble, France.
12. Gross, R., I. Matthews, and S. Baker, *Constructing and Fitting Active Appearance Models With Occlusion.* Proceedings of the IEEE Workshop on Face Processing in Video, June 2004.
13. Ishiyama, R. and S. Sakamoto, *Geodesic illumination basis: Compensating for illumination variations in any pose for face recognition.* International Conference on Pattern Recognition (ICPR), 2002. **4**: p. 297-301.
14. Lyons, M., et al., *Coding facial expressions with gabor wavelets.* Proceedings of International Conference on Face and Gesture Recognition, 1998.
15. Pantic, M. and L.J.M. Rothkrantz, *Expert system for automatic analysis of facial expressions.* Image and Vision Computing 2000. **18**: p. 881-905.
16. Tamminen, T. and J. Lampinen. *A Bayesian Occlusion Model for Sequential Object Matching*. in *British Machine Vision Conference*. 7th-9th Sept, 2004. Kingston, UK.
17. Tian, Y.-l., T. Kanade, and J.F. Cohn, *Recognizing Action Units for Facial Expression Analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. **23**(2): p. 97-116.
18. Cootes, T.F. and C.J. Taylor, *Statistical Models of Appearance for Computer Vision*. 2000, University of Manchester.
19. Turk, M.A. and A.P. Pentland, *Face recognition using eigenfaces.* IEEE Computer Vision and Pattern Recognition, 1991: p. 586-591.
20. Cooper, D.H., et al., *Active shape models - their training and application.* Computer Vision and Image Understanding, 1995: p. 38-59.

21. Cootes, T.F., G.J. Edwards, and C.J. Taylor, *Active appearance models.* Lecture Notes in Computer Science, 1998. **1407**: p. 484–.
22. Cootes, T.F. and E. Taylor. *Active shape models – 'smart snakes'*. in *British Machine Vision Conf, BMVC92*. 1992.
23. Kass, M., A. Witkin, and D. Terzopoulos, *Snakes: Active contour models.* International Journal of Computer Vision, 1987. **1**(4): p. 321-331.
24. Wiskott, L., *Face recognition by elastic bunch graph matching.* IEEE Transactions on Patern Analysis and Machine Intelligence, 1997. **19**(7): p. 775-780.
25. Kanade, T., *Picture processing by computer complex and recognition of human faces.* . 1973, Kyoto Univ., Dept. Inform. Sci.
26. Brunelli, R. and T. Poggio, *Face Recognition: Features versus Templates.* IEEE Trans. on Pattern Analysys and Machine Intelligence, 1993. **15**(10): p. 1042-1052.
27. Ekenel, H.K. and R. Stiefelhagen. *A Generic Face Representation Approach for Local Appearance based Face Verification*. in *Computer Vision and Pattern Recognition Workshop on Face Recognition Grand Challenge Experiments*. June 2005. San Diego, USA.
28. Zhu, Z. and Q. Ji, *Robust Real-Time Eye Detection and Tracking Under Variable Lighting Conditions and Various Face Orientations.* Computer Visage and Image Understanding, 2005. **98**: p. 124-154.
29. Pantic, M., M. Tomc, and L.J.M. Rothkrantz. *A Hybrid approach to mouth features detection*. in *Proceeding of the 2001 Systems, Man and Cybernetics Conference*. 2001.
30. Jollie, I.T., *Principal Component Analysis.* Springer, second edition, October 2002.
31. McInerney, T. and D. Terzopoulos, *Deformable models in medical image analysis: a survey.* Medical Image Analysis, 1996. **1**(2): p. 91-108.
32. Xiao, S.S. and M. Jin, *From 2D to 3D: Using Illumination Cones to Build 3d Face Model.* Journal of Physics: Conference Series, 2006. **48**: p. 318–323.
33. Xu, C., et al. *A new Attempt to Face Recognition Using 3D Eigenfaces*. in *The 6th Asian Conference on Computer Vision (ACCV)*. 2004.
34. Cootes, T.F., et al. *Coupled-View Active Appearance Models*. in *British Machine Vision Conference 2000*. 2000.
35. Pentland, A., B. Moghaddam, and T. Starner. *View-based and modular eigenspaces for face recognition*. in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94)*. 1994. Seattle, WA.
36. Mattews, I., J. Xiao, and S. Baker, *2D vs. 3D deformable face models : Representational power, construction, and real-time fitting.* International journal of computer vision, 2007. **75**(1): p. 93-113.
37. http://www.sic.rma.ac.be/~beumier/DB/3d_rma.html.
38. Shan, T., et al., *Robust Face Recognition Technique for a Real-Time Embedded Face Recognition System.* Pattern Recognition Technologies and Applications: Recent Advances, Idea Group, 2008.
39. Fasel, B. and J. Luettin, *Automatic facial expression analysis: A survey.* Pattern Recognition, 2003.
40. Ekman, P. and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1976.

41. Hager, J., P. Ekman, and W. Friesen, *Facial action coding system.* Salt Lake City, UT: A Human Face, 2002.
42. Ekman, P. and W.V. Friesen, *Constants across cultures in the face and emotion.* Journal of Personality and Social Psychology, 1971. **17**.
43. Rosenblum, M., Y. Yacoob, and L.S. Davis, *Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture*. 1998.
44. Gokturk, S.B., et al., *Model-based face tracking for view independent facial expression recognition.* Proc. IEEE International Conference of Face and Gesture Recognition, 2002: p. 272-278.
45. Chang, Y., et al., *Manifold based analysis of facial expression.* Image & Vision Computing, 2006. **24**(6): p. 605-614.
46. Kotsia, I. and I. Pitas, *Facial expression recognition in image sequences using geometric deformation features and support vector machines.* IEEE Transactions on Image Processing, 2007. **16**(1): p. 172-187.
47. Zhang, Z., et al., *Comparison between geometry-based and Gaborwavelets-based facial expression recognition using multi-layer perceptron.* Third IEEE Intel. Conf. On Automatic Face and Gesture Recognition, 1998.
48. Silapachote, P., D. Karuppiah, and A.R. Hanson, *Feature selection using AdaBoost for face expression recognition*. 2005, Massachusetts University, Amherst department of Computer Science.
49. Gong, S., P.W. McOwan, and C. Shan, *Appearance Manifold of Facial Expression.* Computer Vision in Human-Computer Interaction, Springer, 2005. **3723**: p. 221-230.
50. Bartlett, M.S., et al., *Measuring facial expressions by computer image analysis.* Psychophysiology, 1999. **36**(2): p. 253-263.
51. Wen, Z. and T.S. Huang, *Capturing Subtle Facial Motions in 3D Face Tracking.* Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003. **2**: p. 1343.
52. Zhang, Y. and Q. Ji, *Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005. **27**(5): p. 699-715.
53. Pantic, M. and I. Patras, *Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences.* IEEE Trans. on Systems, Man and Cybernetics - Part B, 2006. **36**(2): p. 433-449.
54. Mase, K., *Recognition of Facial Expression from Optical Flow.* Institut of Electronics Information and Communication Engineers Transactions, 1991. **E74**(10): p. 3474-3483.
55. Liu, Y., et al., *Facial Asymmetry Quantification for Expression Invariant Human Identification.* Computer Vision and Image Understanding Journal 2003. **91**(1/2): p. 138 - 159.
56. Uyl, M.J.d. and H.v. Kuilenburg, *The FaceReader: Online facial expression recognition.* Amsterdam, The Netherlands, http:// www. noldus. com/ site/, 2007.
57. Hong, T., et al., *Facial Expression Recognition Using Active Appearance Model.* Advances in Neural Networks, 2006. **3972**: p. 69-76.
58. Zhang, L., et al., *Robust face alignment based on local texture classifiers.* IEEE International Conference on Image Processing, ICIP, 2005: p. II- 354-7.
59. Shan, S., et al., *Enhanced Active Shape Models with Global Texture Constraints for Image Analysis.* Lecture Notes in Computer Science, 2003. **2871/2003**: p. 593-597.

60. Dailey, M.N. and G.W. Cottrell, *PCA = Gabor for expression recognition.* UCSD CSE TR CS-629, 1999.
61. Black, M.J. and Y. Yacoob, *Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Local Parametric Models of Image Motion.* ICCV, 1995.
62. Sun, W. and Q. Ruan, *Two-Dimension PCA for Facial Expression Recognition.* 8th International Conference on Signal Processing, 2006. **3**: p. 16-20.
63. Pantic, M. and L.J.M. Rothkrantz, *Automatic Analysis of Facial Expressions: The State of the Art.* IEEE Transactions of Pattern Analysis and Machine Intelligence, dec. 2000. **12**(12).
64. Tian, Y., T. Kanade, and J.F. Cohn, *Facial Expression Analysis.* Book Chapter in Handbook of face recognition, S.Z. Li & A.K. Jain, ed., Springer, 2003.
65. M.Pantic and M.S. Bartlett, *Machine Analysis of Facial Expressions.* Machine Analysis of Facial Expressions, 2007: p. 377-416.
66. Y.Tian, T. Kanade, and J.F. Cohn, *Facial Expression Analysis.* Book Chapter in Handbook of face recognition,S.Z. Li & A.K. Jain, ed., Springer, 2003.
67. Pantic, M., *Face for Interface.* The Encyclopedia of Multimedia Technology and Networking, M. Pagani, Ed. Hershy, USA: Idea Group Reference, May 2005. **1**: p. 308-314.
68. Lucey, S., A.B. Ashraf, and J. Cohn, *Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face.* Face Recognition Book, K. Kurihara, ed., Pro Literatur Verlag, Mammendorf, Germany, 2007.
69. Donner, R., et al., *Fast Active Appearance Model Search Using Canonical Correlation Analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006. **28**(10): p. 1690-1694.
70. Yang, M.-H., D.J. Kriegman, and N. Ahuja, *Detecting Faces in Images: A Survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. **24**(1): p. 34-59.
71. http://www.mathworks.com/matlabcentral/fileexchange/19912. [cited.
72. Kim, S., et al. *Robust face recognition using AAM and Gabor features*. in *World Academy of Science, Engineering and Technology*. 2007.
73. Cootes, T., et al. *A unified framework for atlas matching using active appearance models*. in *16th Conference on Information Processing in Medical Imaging*. 1999.
74. Edwards, G., C.J. Taylor, and T.F. Cootes. *Interpreting face images using active appearance models*. in *3rd International Conference on Automatic Face and Gesture Recognition*. 1998. Japan.
75. Edwards, G.J., T.F. Cootes, and C.J. Taylor. *Face Recognition Using Active Appearance Models*. in *5th European Conference on Computer Vision*. 1998.
76. Walker, K.N., T.F. Cootes, and C.J. Taylor. *Determining correspondences for statistical models of facial appearance*. in *Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition*. 2000.
77. Goodall, C., *Procrustes methods in the statistical analysis of shape.* Journal of the Royal Statistical Society B, 1991. **53**(2): p. 285-339.
78. Glasbey, C.A. and K.V. Mardia, *A review of image-warping methods.* Journal of Applied Statistics, 1998. **25**(2): p. 155–172.
79. Stegmann, M.B., *Active Appearance Models: Theory, Extensions and Cases*, in *Informatics and Mathematical Modelling*. 2000, Technical University of Denmark, DTU. p. 262.

80. Stegmann, M.B., *Generative Interpretation of Medical Images*. 2004, Informatics and Mathematical Modelling, Technical University of Denmark.
81. Bookstein, F.L., *Principal warps: Thin-plate splines and the decomposition of deformations.* 1989. **11**(6): p. 567-585.
82. Ionita, M.C., *Advances in the Design of Statistical Face Modelling Techniques for Face Recognition*, in *Department of Electronic Engineering*. 2008, National University of Ireland: Galway.
83. Ionita, M.C., I. Bacivarov, and P. Corcoran. *Separating Directional Lighting Variability in Statistical Face Modeling Based on Texture Space Decomposition*. in *15th International Conference on Digital Signal Processing*. 2007. Cardiff, Wales.
84. Abboud, B. and F. Davoine, *Bilinear factorisation for facial expression analysis and synthesis.* IEE Proceedings of Vision, Image and Signal Processing, 3 June 2005. **152**(3): p. 327- 333.
85. Abboud, B. and F. Davoine. *Appearance factorization for facial expression analysis*. in *British Machine Vision Conference*. 7th-9th Sept, 2004. Kingston, UK.
86. Macedo, I., E.V. Brazil, and L. Velho. *Expression Transfer between Photographs through Multilinear AAM's*. in *SIBGRAPI*. 2006.
87. Edwards, G.J., T.F. Cootes, and C.J. Taylor. *Advances in Active Appearance Models*. in *Seventh IEEE International Conference on Computer Vision*. 1999. Kerkyra, Greece.
88. Cootes, T.F., G. Edwards, and C.J. Taylor, *A comparative evaluation of active appearance model algorithms.* BMVC 98. Proc.of the Ninth British Machine Vision Conf., 1998. **2**: p. 680-689.
89. Sung, J. and D. Kim, *A background robust active appearance model using active contour technique.* Pattern Recognition, 2007. **40**: p. 108 – 120.
90. Jones, M.J. and J.M. Rehg, *Statistical Color Models with Application to Skin Detection.* International Journal of Computer Vision, 2002. **46**(1): p. 81 - 96.
91. Baker, S. and I. Matthews, *Active appearance models revisited.* International Journal of Computer Vision. **60**(2): p. 135-164.
92. Cordea, M.D., E.M. Petriu, and T.E. Whalen. *A 3D-anthropometric-muscle-based active appearance model*. in *IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, (VECIMS)*. 2004.
93. Cristinacce, D. and T.F. Cootes, *Feature detection and tracking with constrained local models.* Proceedings of the British Machine Vision Conference, 2006.
94. Zhang and F.S. Cohen. *Component-based Active Appearance Models for face Modelling*. in *International Conference of Advances in Biometrics, ICB*. January 5-7, 2006. Hong Kong, China.
95. Roberts, M.G., T.F. Cootes, and J.E. Adams. *Linking Sequences of Active Appearance Sub-Models via Constraints: an Application in Automated Vertebral Morphometry*. in *British Machine Vision Conference*. 2003.
96. Peyras, J., et al. *Segmented AAMs Improve Person-Independent Face Fitting*. in *Eighteenth British Machine Vision Conference*. 2007. Warwick, UK.
97. Zalewski, L. and S. Gong. *2D statistical models of facial expressions for realistic 3D avatar animation*. in *Computer Vision and Pattern Recognition, CVPR*. 20-25 June 2005.

98. Doretto, G. and S. Soatto, *Dynamic Shape and Appearance Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006. **28**(12): p. 2006-2019.
99. Zhou, S.K., et al., *Pairwise Active Appearance Model and Its Application to Echocardiography Tracking.* Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Springer, 2006. **4190/2006**: p. 736-743.
100. Wang, W., et al. *Combining Active Shape Models and Active Appearance Models For Accurate Image Interpretation*. in *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2003. Taipei, Taiwan.
101. Liu, X. *Generic face alignment using boosted appearance model*. in *IEEE Conference on Computer Vision and Pattern Recognition*. 2007.
102. Cootes, T.F., G.J. Edwards, and C.J. Taylor., *Active appearance models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. **23**(6): p. 681-685.
103. Hou, X., et al., *Direct appearance models.* Computer Vision and Pattern Recognition, 2001: p. 828-833.
104. Batur, A.U. and M.H. Hayes, *Adaptive active appearance models.* IEEE Transactions on Image Processing, 2005. **14**(11): p. 1707 - 1721.
105. Matthews, I. and S. Baker, *Active appearance models revisited.* International Journal of Computer Vision, 2004. **60**(2): p. 135-164.
106. Shrum, S.K. *Relationship between gender and personality characteristics assigned to various facial expressions*. in *Loyola University*. 2001.
107. Saatci, Y. and C. Town, *Cascaded Classification of Gender and Facial Expression using Active Appearance Models.* 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006.
108. Kanade, T., J.F. Cohn, and T. Yingli, *Comprehensive database for facial expression analysis.* Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000: p. 46 - 53.
109. Yuki, M., W.W. Maddux, and T. Masuda, *Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States.* Journal of Experimental Social Psychology, 2007. **43**: p. 303-311.
110. Elfenbein, H.A. and N. Ambady, *When Familiarity Breeds Accuracy: Cultural Exposure and Facial Emotion Recognition.* Journal of Personality and Social Psychology, 2003. **85**(2): p. 276–290.
111. Sackeim, H.A., R.C. Gur, and M.C. Saucy, *Emotions are expressed more intensely on the left side of the face* Science, 1978. **202**(4366): p. 434 - 436.
112. Mendolia, M. and R.E. Kleck, *Watching people talk about their emotions—Inferences in response to full-face vs. profile expressions.* Motivation and Emotion 1991. **15**(4): p. 229-242.
113. Hess, U., S. Blairy, and R.E. Kleck, *The Influence of Facial Emotion Displays, Gender, and Ethnicity on Judgments of Dominance and Affiliation.* Journal of Nonverbal Behavior, 2000. **24**: p. 265-283.
114. Tamminen, T., et al. *Joint Modelling of Facial Expression and Shape from Video, Image Analysis*. in *14th Scandinavian Conference, SCIA* 2005. Joensuu, Finland: Lecture Notes in Computer Science.
115. Beszedes, M. and P. Culverhouse, *Facial Emotions and Emotion Intensity Levels Classification and Classification Evaluation.* Proceedings of British Machine Vision Conference 2007.

116. Lee, H. and D. Kim, *Facial Expression Transformations for Expression-Invariant Face Recognition.* Advances in Visual Computing, Springer, 2006. **4291**.
117. Datcu, D. and L. Rothkrantz. *Facial Expression Recognition in still pictures and videos using Active Appearance Models. A comparison approach*. in *Proceedings of the 2007 international conference on Computer systems and technologies*. 2007. Bulgaria.
118. Bartlett, M.S., et al., *Automatic Recognition of Facial Actions in Spontaneous Expressions.* Multimedia, September 2006. **1**(6): p. 22-35.
119. Costen, N., et al. *Automatic extraction of the face identity-subspace*. in *British Machine Vision Conference*. 1999.
120. Buenaposada, J.M., E. Munoz, and L. Baumela. *Efficiently estimating facial expression and illumination in appearance-based tracking*. in *BMVC*. September 2006. Edimburgh, UK.
121. Mercier, H. and P. Dalle, *Face analysis : identity vs. Expressions.*
122. Xin, S. and H. Ai, *Face Alignment under Various Poses and Expressions.* Affective Computing and Intelligent Interaction, Springer, 2005. **3784**.
123. Marshall, D., et al. *Speech and Expression Driven Animation of a Video-Realistic Appearance Based Hierarchical Facial Model*. in *Workshop in conjunction with IEEE CVPR of Learning, Representation and Context for Human Sensing in Video*. June 22nd 2006. New York, USA.
124. Li, X., G. Mori, and H.Zhang. *Expression-Invariant Face Recognition with Expression Classification*. in *The 3rd Canadian Conference on Computer and Robot Vision*. 2006.
125. Antonini, G., et al., *Discrete Choice Models for Static Facial Expression Recognition.* Advanced Concepts for Intelligent Vision Systems, 2006. **4179**: p. 710-721.
126. Viola, P. and M.J. Jones, *Robust Real-Time Face Detection.* International Journal of Computer Vision, 2004. **57**(2): p. 137–154.
127. Dornaika, F. and J. Ahlberg. *Face Model Adaptation for Tracking and Active Appearance Model Training*. in *14th British Machine Vision Conference (BMVC)*. September 2003. Norwich, UK.
128. Cascia, M.L., S. Sclaroff, and V. Athitsos, *Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, April 2000. **22**(4).
129. Dornaika, F. and J. Ahlberg, *Fast and reliable active appearance model search for 3d face tracking.*
130. Gonzalez, J., et al. *Bilinear Active Appearance Models*. in *Workshop on Non-rigid Registration and Tracking through Learning*. 2007.
131. Abboud, B., F. Davoine, and M. Dang, *Facial expression recognition and synthesis based on an appearance model.* Signal Processing: Image Communication, 2004. **19**(8).
132. Abboud, B., F. Davoine, and M. Dang. *Expressive face recognition and synthesis*. in *IEEE workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*. June 2003. Madison, U.S.A.
133. Butakoff, C. and A.F. Frangi, *A Framework for Weighted Fusion of Multiple Statistical Models of Shape and Appearance.* IEEE Transactions on Analysis and Machine Intelligence, 2006. **28**(11).

134. Chuang, E.S., H. Deshpande, and C. Bregler, *Facial Expression Space Learning*. Pacific Graphics, 2002.
135. Ahlberg, J. *Using the Active Appearance Algorithm for Face and Facial Feature Tracking*. in *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*. 2001.
136. Ahlberg, J. and R. Forchheimer, *Face Tracking for Model-based Coding and Face Animation.* International Journal of Imaging Systems and Technology (IJIST), 2003. **13**(1): p. 8-22.
137. Chen, Y. and F. Davoine. *Simultaneous tracking of rigid head motion and non-rigid facial animation by analyzing local features statistically*. in *British Machine Vision Conference*. 4-7 September 2006. Edinburgh , UK.
138. Dedeoglu, G., T. Kanade, and S. Baker, *The Asymmetry of Image Registration and Its Application to Face Tracking.* IEEE Transactions on Pattern Analysis and machine Intelligence, 2007. **29**(5).
139. Edwards, G.J., C.J. Taylor, and T.F. Cootes. *Learning to Identify and Track Faces in Image Sequences*. in *Sixth International Conference on Computer Vision (ICCV'98)*. 1998.
140. Kotsia, I., et al. *Texture and Shape Information Fusion for Facial Action Unit Recognition*. in *First International Conference on Advances in Computer-Human Interaction (ACHI)*. 2008.
141. Kotsia, I., S. Zafeiriou, and I. Pitas, *Texture and shape information fusion for facial expression and facial action unit recognition.* Pattern Recognition, 2008. **41**: p. 833 – 851.
142. Lee, C. and A. Elgammal. *Nonlinear Shape and Appearance Models for Facial Expression Analysis and Synthesis*. in *18th International Conference on Pattern Recognition*. 2006. Washington, DC.
143. Torre, F.D.l. and M. Black, *Robust parameterized component analysis: theory and applications to 2D facial appearance models.* Computer Vision and Image Understanding 2003. **91**: p. 53–71.
144. Dornaika, F. and J. Ahlberg. *Efficient Active Appearance Model for Real-Time Head and Facial Feature Tracking*. in *IEEE International Workshop on Analysis and Modelling of Faces and Gestures (AMFG)* 17 Oct. 2003.
145. Choi, H.-C. and S.-Y. Oh. *Realtime Facial Expression Recognition using Active Appearance Model and Multilayer Perceptron*. in *SICE-ICASE International Joint Conference*. Oct. 18-2 1 2006. Bexco, Busan, Korea.
146. Theobald, B., et al. *Real-time expression cloning using active appearance models*. in *ACM International Conference on Multimodal Interfaces (ICMI'07)*. 2007.
147. Choi, H.C. and S.Y. Oh. *Real-time Recognition of Facial Expression using Active Appearance Model with Second Order Minimization and Neural Network*. in *International IEEE Conference on Systems, Man and Cybernetics, SMC 06*. Oct. 2006.
148. Cho, K.S. and Y.G. Kim, *Continuous Recognition of Human Facial Expressions Using Active Appearance Model.* Universal Access in Human-Computer Interaction, Ambient Interaction, Springer, 2007: p. 777-783.
149. Cho, K.S., Y.G. Kim, and Y.B. Lee. *Real-time Expression recognition System using Active Appearance Model and EFM*. in *International Conference on Computational Intelligence and Security*. Nov. 2006.

150. Li, H., H. Lin, and G. Yang. *A New Facial Expression Analysis System Based on Warp Image*. in *6th World Congress on Intelligent Control and Automation*. June 21 - 23, 2006. Dalian, China.
151. Bartlett, M.S., J.R. Movellan, and T.J. Sejnowski, *Face Modeling by Information Maximization.* Face Processing: Advanced Modeling and Methods, Elsevier: p. 219-253.
152. Ratliff, M.S. and E. Patterson, *Emotion Recognition using Facial Expressions with Active Appearance Models.* Proceeding Human Computer Interaction – ACTA Press, 2008.
153. Sorci, M., G. Antonini, and J. Thiran, *Relevant Component Analysis for static facial expression classification*. 2005, EPFL-ITS.
154. Oliver, N., A. Pentland, and F. Berard, *LAFTER: Lips and Face Real Time Tracker with Facial Expression Recognition.* IEEE Conference of Computer Vision and Pattern Recognition, 1997.
155. Sebe, N., et al. *Emotion recognition using a Cauchy naive Bayes classifier*. in *16th International Conference on Pattern Recognition (ICPR 2002), 11-15 August 2002, Quebec, Canada*.
156. Vishnubhotla, S., *Support Vector Classification*. 2005.
157. Dasarathy, B.V., *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. 1991.
158. Boser, B., I. Guyon, and V. Vapnik, *A training algorithm for optimal margin classifiers.* Fifth Annual Workshop on Computational Learning Theory. ACM Press, Pittsburgh, 1992.
159. Ahmad, A.R., et al., *The Comparative Study of SVM Tools for Data Classification.*
160. Wilhelm, T., H.-J. Bohme, and H.-M. Gross. *Classification of Face Images for Gender, Age, Facial Expression, and Identity*. in *Artificial Neural Networks (ICANN'05)*. 2005. Warsaw, LNCS: Springer Verlag.
161. Kuilenburg, H., M. Wiering, and M.d. Uyl. *A Model Based Method for Automatic Facial Expression Recognition*. in *European Conference on Machine Learning*. 2005.
162. Liebelt, J., J. Xiao, and J. Yang, *Robust AAM Fitting by Fusion of Images and Disparity Data.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. **2**: p. 2483-2490.
163. Melgani, F. and L. Bruzzone, *Classification of hyperspectral remote sensing images with Support Vector Machines.* IEEE Transactions on Geoscience and Remote Sensing, 2004. **42**: p. 1778 – 1790.
164. Gualtieri, J.A. and R.F. Cromp. *Support vector machines for hyperspectral remote sensing classification*. in *27th AIPR Workshop: Advances in Computer Assisted Recognition*. 1998. Washington, DC: SPIE.
165. Phillips, P.J., et al., *The FERET Evaluation Methodology for Face Recognition Algorithms.* IEEE Trans. Pattern Analysis and Machine Intelligence, 2000. **22**: p. 1090-1104.
166. Phillips, P.J., et al., *The FERET database and evaluation procedure for face recognition algorithms.* Image and Vision Computing, 1998. **16**(5): p. 295-306.
167. Pantic, M., et al. *Web-based database for facial expression analysis*. in *IEEE International Conference on Multimedia and Expo (ICME'05), http://www.mmifacedb.com*. 2005.

168. Ekman, P., *Emotions Revealed: Recognizing Faces and Feeling to Improve Communication and Emotional Life.* 2003.
169. Wallhoff, F., *The Facial Expressions and Emotions Database Homepage (FEEDTUM), www.mmk.ei.tum.de/~waf/fgnet/feedtum.html.* Sept. 2005.
170. Wallhoff, F., et al. *Efficient Recognition of authentic dynamic facial expressions on the FEEDTUM database*. in *IEEE International Conference on Multimedia and Expo*. 9-12 July 2006.
171. *Japanese Female Facial Expression (JAFFE) Database, www.mic.atr.co.jp/~mlyons/jaffe.html.*
172. Gunes, H. and M. Piccardi, *A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior*, in *the 18th International Conference on Pattern Recognition (ICPR)*: Hong Kong, China.
173. Lundqvist, D., A. Flykt, and A. Öhman, *The Karolinska Directed Emotional Faces.* Psychology section, Karolinska Institute, 1998.
174. Loh, M., Y. Wong, and C. Wong. *Facial Expression Recognition for E-learning Systems using Gabor Wavelet & Neural Network.* in *Sixth International Conference on Advanced Learning Technologies (ICALT'06)*. 2006.
175. O'toole, A.J., K.A. Deffenbacher, and K.M. D. Valentin, D. Huff, *The perception of face gender: The role of stimulus structure in recognition and classification.* Memory and Cognition, 1998. **26**: p. 146-160.
176. *Yale database, http://cvc.yale.edu*.
177. A. Martinez, R.B., *The AR Face Database*. 1998, Computer Vision Centre, Barcelona.
178. *http://www.paulekman.com/researchproducts.html*.
179. Solina, F., et al. *Colour-based face detection in the 15 seconds of fame art installation*. in *Computer Vision / Computer Graphics Collaboration for Model-based Imaging, Rendering, image Analysis and Graphical special Effects*. 2003. INRIA Rocquencourt, France.
180. *http://www.fgnet.rsunit.com/index.php?page=specifications.*
181. Ekman, P. and W.V. Friesen, *Pictures of Facial Affect.* Palo Alto: Consulting Psychologists Press, 1976.
182. *http://facedb.blogspot.com/*.
183. Beaupré, M.G. and U. Hess, *Cross-cultural emotion recognition among Canadian ethnic groups.* Cross-cultural Psychology, 2005. **36**(3): p. 355-370.
184. Rose., I., *The anatomy of the Human Eye*, IvyRose Ltd., Hampden House, Monument Business Park, Warpsgrove Lane, Chalgrove, Oxford (February 2006). Available on http://www.ivyrose.co.uk.
185. Ji, Q. and X. Yang. *Real Time Visual Cues Extraction for Monitoring Driver Vigilance*. in *ICVS '01, Second International Workshop on Computer Vision Systems*. 2001.
186. Orozco, J., et al. *Human Emotion Evaluation on Image Sequences*. in *CogSys II Conference*. 2006. Radboud University Nijmegen, Netherlands.
187. Karson, C.N., et al., *Blink rates and disorders of movement.* Neurology, 1984. **34**: p. 677-678.
188. Karson., C.N., *Physiology of normal and abnormal blinking.* Advances in Neurology, 1988. **49**: p. 25-37.
189. Brandreth, G., *Your Vital Statistics.* p. 26.
190. http://www.theregister.co.uk/2006/06/30/the_odd_body_blinking/. [cited.

191. Stern, J.A., D. Boyer, and D.J. Schroeder, *Blink rate as a measure of fatigue: A review (Final Report DOT/FAA/AM-94/17) Washington, DC: Federal Aviation Administration.* 2004.
192. Stern, J.A., et al., *Blinks, saccades, and fixation pauses during vigilance task performance: I. Time on task*. 1994, National Technical Information Service: Springfield, Virginia.
193. Morimoto, C.H., et al., *Pupil detection and tracking using multiple light sources.* Image and Vision Computing, 2000. **18**(4): p. 331-335.
194. Lee, S.P., J.B. Badler, and N.I. Badler, *Eyes alive.* ACM Trans. Graph, 2002. **21**(3): p. 637-644.
195. Tekalp, A.M. and J. Ostermann, *Face and 2-D mesh animation in MPEG-4.* Signal Processing: Image Communication, 2000. **3**: p. 387-421.
196. *Ssontech SynthEyes: http://www.ssontech.com/.*
197. Betke, M., J. Gips, and P. Fleming, *The Camera Mouse: Visual tracking of body features to provide computer access for people with severe disabilities.* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2002.
198. Chau, M. and M. Betke, *Real Time Eye Tracking and Blink Detection with USB Cameras*. 2005, Boston University Computer Science.
199. Moriyama, T., et al. *Automatic recognition of eye blinking in spontaneously occurring behaviour*. in *16th International Conference of Pattern Recognition*. 2002.
200. Gorodnichy, D.O. *Second Order Change Detection, and its Application to Blink-Controlled Perceptual Interfaces*. in *International Association of Science and Technology for Development (IASTED)*. 2003. Benalmadena, Spain.
201. Kawato, S. and N. Tetsutani. *Detection and tracking of eyes for gaze-camera control*. in *Intern. Conf. on Vision Interface (VI'2002)*. 2002. Calgary.
202. Grauman, K., et al. *Communication via Eye Blinks- Detection and Duration Analysis in Real Time*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2001. Lihue, HI.
203. Tan, H., Y.-J. Zhan, and R. Li. *Robust eye extraction using deformable template and feature tracking ability*. in *Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing*. 2003.
204. Hansen, D.W. and A.E.C. Pece, *Eye tracking in the wild.* Computer Vision and Image Understanding, 2005. **98**: p. 155–181.
205. Yuille, A., P. Hallinan, and D. Cohen, *Feature extraction from faces using deformable templates.* International Journal of Computer Vision, 1992. **8**(2): p. 99-111.
206. Kawato, S. and J. Ohya. *Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes"*. in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*. 2000.
207. Kothari, R. and J.L. Mitchell. *Detection of eye locations in unconstrained visual images*. in *International Conference on Image Processing*. 1996.
208. Hansen, D.W., et al. *Eye Typing using Markov and Active Appearance Models*. in *Sixth IEEE Workshop on Applications of Computer Vision (WACV)*. 2002.
209. Lam, K. and H. Yan, *Locating and extracting the eye in human face images.* Pattern Recognition, 1996. **29**(5): p. 771-779.

210. Moriyama, T., et al., *Meticulously Detailed Eye Region Model and Its Application to Analysis of Facial Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006. **28**(5).
211. *Represent Eye's Appearance by Numbers (Automated individualization of the eye region model using AAM).* http: //www.ozawa.ics.keio.ac.jp/ ~moriyama /projects/esa-aam.htm.
212. Tian, Y., T. Kanade, and J.F. Cohn. *Dual-state Parametric Eye Tracking*. in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*. 2000.
213. Pregenzer, M. and G. Pfurtscheller, *Frequency component selection for an EEG-based brain to computer interface.* IEEE Trans. Rehab. Eng, 1999. **7**(4): p. 413–419.
214. DiMattia, et al., *An Eye Control Teaching Device for Students Without Language Expressive Capacity: EagleEyes.* U.K.: Edwin Mellen, 2001.
215. Glenstrup, A. and T. Angell-Nielsen, *Eye controlled media: Present and future state*, in *Laboratory of Psychology*. 1995, University of Copenhagen.
216. Zhu, Z. and Q. Ji, *Eye and gaze tracking for interactive graphic display*. Machine Vision and Applications, Springer, 2004.
217. Hansen, D. and A. Pece. *Iris tracking with feature free contours*. in *Workshop on Analysis and Modelling of Faces and Gestures (AMFG)*. 2003.
218. Magee, J.J., et al., *EyeKeys: A Real-time Vision Interface Based on Gaze Detection from a Low-grade Video Camera.* IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2004.
219. Corcoran, P., M. Ionita, and I. Bacivarov. *Next Generation Face Tracking Technology Using AAM Techniques*. in *Signals, Circuits and Systems, ISSCS 2007, International Symposium*. 2007. Iasi, Romania.
220. Pantic, M. and L.J.M. Rothkrantz, *Automatic Analysis of Facial Expressions: The State of the Art.* IEEE Transactions of Pattern Analysis and Machine Intelligence, 2000. **12**(12).
221. Pahor, V. and S. Carrato. *A Fuzzy Approach to Mouth Corner Detection*. in *Proceedings of the 1999 International Conference on Image Processing (ICIP '99)*. 1999. Kobe, Japan.
222. Liew, A.W.-C., S.H. Leung, and W.H. Lau, *Segmentation of Colour Lip Images by Spatial Fuzzy Clustering.* IEEE Transactions on Fuzzy Systems, 2003. **11**(4): p. 542 – 549.
223. Lucey, S., S. Sridharan, and V. Chandran, *Adaptive mouth segmentation using chromatic features.* Pattern Recognition Letters, 2002. **23**: p. 1293–1302.
224. Chindaro, S. and F. Deravi. *Directional Properties of Colour Co-occurrence Features for Lip Location and Segmentation*. in *The 3rd International Conference on Audio and Video-Based Biometric Person Authentication*. 2001.
225. Hennecke, M., V. Prasad, and D. Stork. *Using deformable templates to infer visual speech dynamics*. in *287h Annual Asimolar Conference on Signals, Systems, and Computer*. 1994. Pacific Grove.
226. Liew, A.W.C., S.H. Hung, and W.H. Lau. *Lip contour extraction using a deformable model*. in *Proceedings of the International Conference on Image Processing*. 2000.
227. Gacon, P., P. Coulon, and G. Bailly. *Non-linear Active Model for mouth inner and outer contours detection*. in *the 13th European Signal Processing Conference - EUSIPCO*. 2005. Antalya, Turquey.

228. Jang, K.S., *Lip Contour Extraction based on Active Shape Model and Snakes.* IJCSNS International Journal of Computer Science and Network Security, 2007. **7**(10).
229. Bacivarov, I., M. Ionita, and P. Corcoran, *Statistical Models of Appearance for Eye Tracking and Eye-Blink Detection and Measurement.* IEEE Transactions on Consumer Electronics, 2008.
230. Foely, J., et al., *Computer Graphics: Principals and Practice, Addison Wesley, Reading, MA*. 1990.
231. Kang, L., et al. *Robust AAM building for morphing in an image-based facial animation system*. in *2008 IEEE International Conference on Multimedia and Expo*. 2008.
232. Matthews, I., et al., *Extraction of Visual Features for Lipreading.* IEEE Transactions on Patern Analysis and Machine Intelligence, February 2002. **24**(2): p. 198-213.
233. Bacivarov, I., M. Ionita, and P. Corcoran, *Statistical Models of Appearance for Eye Tracking and Eye-Blink Detection and Measurement.* IEEE Transactions on Consumer Electronics, August 2008. **54**(3): p. 1312-1320
234. Bacivarov, I., M.C. Ionita, and P. Corcoran. *A Combined Approach to Feature Extraction for Mouth Characterization and Tracking*. in *ISSC*. 2008. Galway, Ireland.
235. Corcoran, P., I. Bacivarov, and M.C. Ionita. *A Statistical Modeling based System for Blink Detection in Digital Cameras*. in *ICCE*. 2008. Las Vegas, USA.
236. Eveno, N., A. Caplier, and P.Y. Coulon. *New color transformation for lips segmentation*. in *IEEE Fourth Workshop on Multimedia Signal Processing*. 2001.
237. Stegmann, M.B., B.K. Ersbøll, and R. Larsen, *FAME - a flexibleappearance modelling environment.* Available: http://www2.imm.dtu.dk/pubdb/p.php?1918, IEEE Transactions on Medical Imaging, 2003. **22**(10): p. 1319–1331.
238. Georgia Tech Face Database, *ftp://ftp.ee.gatech. edu/pub/users/hayes/facedb/.*
239. Sanderson, C. and K.K. Paliwal, *Identity Verification Using Speech and Face Information.* Digital Signal Processing, 2004. **14**(5): p. 449-480.
240. Thompson, P., *A new illusion.* Perception, 1980. **9**: p. 483–484.
241. Gross, R., I. Matthews, and S. Baker, *Generic vs. person specific active appearance models.* Image and Vision Computing, 2005. **23**(11): p. 1080-1093.
242. Nusseck, M., et al., *The contribution of different facial regions to the recognition of conversational expressions.* Journal of Vision, 2008. **8**(8): p. 1-23.
243. Kwon, Y.H. and N.d.V. Loboy, *Age Classification from Facial Images.* Computer Vision and Image Understanding, 1999. **74**(1): p. 1-21.
244. Ou, Y., et al. *A Real Time Race Classification System*. in *IEEE International Conference on Information Acquisition*. 2005. Hong Kong and Macau, China.
245. Meynet, J., *Information theoretic combination of classifiers with application to face detection*. 2007, Ecole polytechnique fédérale de Lausanne EPFL.
246. Freund, Y., Y. Mansour, and R. Schapire. *Why averaging classifiers can protect against overfitting*. in *the Eighth International Workshop on Artificial Intelligence and Statistics*. 2001.

247. Smith, L.I., *A tutorial on Principal Components Analysis*. http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, 2002.
248. Ohta, Y., T. Kanade, and T. Sakai, *Color information for region segmentation.* Computer Graphics and Image Processing, 1980. **13**: p. 222-241.