

Statistical Modelling for Facial Expression Dynamics

Lukasz Zalewski

Submitted to the University of London in partial fulfilment of the requirements for
the degree of Doctor of Philosophy

Queen Mary, University of London

2012

Statistical Modelling for Facial Expression Dynamics

Lukasz Zalewski

Abstract

One of the most powerful and fastest means of relaying emotions between humans are facial expressions. The ability to capture, understand and mimic those emotions and their underlying dynamics in the synthetic counterpart is a challenging task because of the complexity of human emotions, different ways of conveying them, non-linearities caused by facial feature and head motion, and the ever critical eye of the viewer. This thesis sets out to address some of the limitations of existing techniques by investigating three components of expression modelling and parameterisation framework: (1) Feature and expression manifold representation, (2) Pose estimation, and (3) Expression dynamics modelling and their parameterisation for the purpose of driving a synthetic head avatar.

First, we introduce a hierarchical representation based on the Point Distribution Model (PDM). Holistic representations imply that non-linearities caused by the motion of facial features, and intra-feature correlations are implicitly embedded and hence have to be accounted for in the resulting expression space. Also such representations require large training datasets to account for all possible variations. To address those shortcomings, and to provide a basis for learning more subtle, localised variations, our representation consists of tree-like structure where a holistic root component is decomposed into leaves containing the jaw outline, each of the eye and eyebrows and the mouth. Each of the hierarchical components is modelled according to its intrinsic functionality, rather than the final, holistic expression label.

Secondly, we introduce a statistical approach for capturing an underlying low-dimension expression manifold by utilising components of the previously defined hierarchical representation. As Principal Component Analysis (PCA) based approaches cannot reliably capture variations caused by large facial feature changes because of its linear nature, the underlying dynamics manifold for each of the hierarchical components is modelled using a Hierarchical Latent Variable Model (HLVM) approach. Whilst retaining PCA properties, such a model introduces a probability density model which can deal with missing or incomplete data and allows discovery of internal within cluster structures. All of the model parameters and underlying density model are automatically estimated during the training stage. We investigate the usefulness of such a model to larger and unseen datasets.

Thirdly, we extend the concept of HLVM model to pose estimation to address the non-linear shape deformations and definition of the plausible pose space caused by large head motion. Since our head rarely stays still, and its movements are intrinsically connected with the way we perceive and understand the expressions, pose information is an integral part of their dynamics. The proposed

approach integrates into our existing hierarchical representation model. It is learned using sparse and discretely sampled training dataset, and generalises to a larger and continuous view-sphere.

Finally, we introduce a framework that models and extracts expression dynamics. In existing frameworks, explicit definition of expression intensity and pose information, is often overlooked, although usually implicitly embedded in the underlying representation. We investigate modelling of the expression dynamics based on use of static information only, and focus on its sufficiency for the task at hand. We compare a rule-based method that utilises the existing latent structure and provides a fusion of different components with holistic and Bayesian Network (BN) approaches. An Active Appearance Model (AAM) based tracker is used to extract relevant information from input sequences. Such information is subsequently used to define the parametric structure of the underlying expression dynamics. We demonstrate that such information can be utilised to animate a synthetic head avatar.

Submitted to the University of London in partial fulfilment of the requirements for
the degree of Doctor of Philosophy

Queen Mary, University of London

2012

Acknowledgements

I would like to thank my supervisor Professor Shaogang Gong for introducing me to computer vision, for his invaluable guidance, patience and enthusiastic supervision throughout the course of this work. I do not think I could have asked for a better supervisor.

I am grateful to Andrew Graves, Richard Howarth, David Russell, Ioanis Tziakos and Prathap Nair for comments and proof-reading the thesis.

I would also like to thank members of the department with whom i worked and socialised, in particular Dennis Parkinson, Tao Xiang, Lourdes Agapito, Peter McOwan, Fabrizio Smeraldi, Christof Monz, Norman Fenton, Martin Neil, Thomas Roelleke, Mounia Lalmas, William Marsh, Jane Reid, Tassos Tombros, Pat Healey, Edmund Robinson, Jon Rowson, Bob Koger, Paulo Oliva, Jamie Sherah, Yogesh Raja, Daniel Garcia, Hayley Hung, Jian Li, Caifeng Shan, Samuel Pachoud, Chen Change Loy, Alex Leung, Caifeng Shan, Chris Jia Kui, Xavier Llado, Bryan Prosser, Andy Anderson, Keith Anderson, Adam Sherwood, Louise Valgerdur Nickerson, Jose Galan, Dino Distefano, Marco Paladini, Milan Verma, Colombine Gardair, Arash Eshghi, Chrystie Mykietiak, Matteo Bregonzio, Alessio Del Bue, Vejen Hlebanov, Jeffrey Ng Sing Kwong, Jun Li, Wang Yong, Jae Ho Lee, Melanie Aurnhammer, Gabriella Kazai, Theodora Tsikrika, George Papatzanis, Oussama Metatla and Hany Azzam.

My thanks also go to the technical staff - Keith Clarke, Matt Bernstein, Simon Boggis, Derek Coppen, David Hawes, Tim Kay, Thomas King and Colin Powell for solving various software and hardware issues that surfaced during the course of this work, and to administrative staff - Joan Hunter, Melissa Yeo, Gill Carter, Sue White, Carla Benjamin, Julie McDonald, Rupal Vaja, Karen Finesilver and Carly Wheeler for their help and advise throughout the years.

Most of all, I am indebted to my family and friends for their encouragement, endless sacrifice and

support. I dedicate this work to them.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged.

Some parts of the work have previously been published as:

1. L. Zalewski and S. Gong. A probabilistic hierarchical framework for expression classification. In *Symposium on Language, Speech and Gesture for Expressive Characters, AISB Convention*, pages 12-20, Leeds, 2004.
2. L. Zalewski and S. Gong. Synthesising and recognition of facial expressions in virtual 3D views. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 493-498, Seoul, 2004.
3. L. Zalewski and S. Gong. A statistical virtual head animator. In *IEE International Conference on Visual Information Engineering*, pages 31-38, Glasgow, 2005.
4. L. Zalewski and S. Gong. 2D statistical models of facial expressions for realistic 3D avatar animation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 217-222, San Diego, 2005.

Lukasz Zalewski

July, 2011

Contents

1	Introduction	18
1.1	Approach	21
1.1.1	Feature Representation	21
1.1.2	Pose Estimation	22
1.1.3	Dynamics Extraction and Parameterisation	23
1.2	Contributions	24
1.3	Overview of the Thesis	24
2	Literature Review	26
2.1	Feature Representation	27
2.1.1	Conceptual Level Representation	28
2.1.2	Data Level Representation	29
2.1.3	Expression Manifold Representation	36
2.2	Semantics Formulation	38
2.2.1	Pose Information	40
2.2.2	Expression Categories Extraction	42
2.3	Summary	49
3	Hierarchical Feature Representation	51
3.1	Active Appearance Models	52
3.1.1	Shape Component	52
3.1.2	Appearance Component	55

3.1.3	Shape and Appearance Combined	57
3.2	Expression Manifold Representation	59
3.2.1	Hierarchical Representation	60
3.3	Hierarchical Latent Variable Model	64
3.4	Experiment	68
3.4.1	Eye Models	69
3.4.2	Mouth Model	75
3.5	Discussion	77
4	Pose Estimation	80
4.1	Pose Model	81
4.2	Experiment	86
4.3	Synthesis	95
4.4	Improved AAM Fitting	99
4.5	Discussion	102
5	Modelling Expression Dynamics	104
5.1	Expression Parameterisation	105
5.1.1	Framework Overview	106
5.1.2	Rule-based fusion	108
5.1.3	Severity Criterion	109
5.2	3D Animation Model	111
5.3	Experiment	112
5.4	Discussion	117
6	Conclusions and Future Work	122
6.1	Conclusions	122
6.2	Future Work	124

List of Figures

1.1	Gollum character in Lord of the Rings.	20
2.1	The 90 feature points used for PDM representation. From Huang and Huang [53].	29
2.2	(a) Location features, where $L1, L2, L3, L4, L5$ correspond to distances between six facial features. (b) Shape features with normalized image on the left and zones of the edge map on the right. From Tian et al. [105].	30
2.3	(a) 20 frontal fiducial facial points (from Valstar and Pantic [110]) (b) 19 frontal and 10 profile fiducial facial points (from Pantic and Rothkrantz [89]).	31
2.4	Motion Units (MUs) representation. From Cohen et al. [21].	32
2.5	Facial features resulting from Gabor wavelet convolution. On the left are two facial expressions, and on the right Gabor representations representing four differently oriented kernels. From Fasel and Luetttin [41].	33
2.6	34 fiducial points representing facial geometry. From Zhang et al. [129].	35
2.7	Universal expressions (from left to right): Fear, Joy, Disgust, Surprise, Sadness, Anger. From Kanade et al. [56].	39
2.8	Some Action Units corresponding to the upper part of the face. From Tian et al. [104].	40
2.9	The 5-point intensity scale of FACS. From Bartlett et al. [6].	43
2.10	Visualisation of Naive Bayes (a) and Tree Augmented Naive Bayes (b) Network of Cohen et al. [21].	45
2.11	Multi level HMM model. From Cohen et al. [21].	47
3.1	Selected training samples with overlaid PDM mask.	53

3.2	Effects of varying the first three (top to bottom) shape parameters in turn between ± 2.5 standard deviation.	55
3.3	Two selected training images (top and bottom) with corresponding shape mask (left) and shape free texture patch (right)	56
3.4	Effects of varying the first three texture parameters in turn between ± 2.5 standard deviations.	57
3.5	Shape variation obtained by varying the first three appearance parameters in turn between ± 2.5 standard deviations.	58
3.6	First (left column) and second (right column) are the two principal components of the training set that have been projected onto shape (top row), texture (middle) and combined shape and texture (bottom) in the expression space.	61
3.7	Original image (top rows in (a) and (b)), and the regions of the image exhibiting motion during expressions (bottom rows) such as surprise (a) and grin (b).	62
3.8	Structure of our hierarchy. The top row corresponds to the highest point in the hierarchy (root), the middle row corresponds to the leaves.	62
3.9	Example expressions that would have to be included in the training set if holistic model was used. In case of hierarchical model only two are necessary.	63
3.10	Conceptual visualisation of the Hierarchical Latent Variable Model. The top level corresponds to a single latent variable model, and subsequent levels correspond to fine-grained mixtures of them.	65
3.11	Sample of correctly (top) and incorrectly (bottom) tracked expressions.	69
3.12	Visualisation of the hierarchical clustering in the eye space for (a) left and (b) right eye models. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy. The numbered clusters in the top level correspond to the order of the cluster frames in the lower levels.	70
3.13	Hierarchical clustering in the eye space. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy.	72

3.14	Classification results for test sequence T1 (a) and test sequence T2 (b) using separate eye models for each of the eyes.	74
3.15	Classification results for test sequence T1 (a) and test sequence T2 (b) using the single eye model.	74
3.16	Hierarchical clustering in the mouth space. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy. . . .	76
3.17	Classification results for the mouth model.	77
4.1	Invalid shape reconstructions of profile views (b). These were generated using modes of variations of a linear shape model trained at near frontal views (a).	81
4.2	The shape variation of facial expression images from $[-40^\circ, 40^\circ]$ 3D views (in yaw) projected onto the 1st three principal components. The manifold forms continuous and separable clusters: $[-40^\circ, -20^\circ]$ (shown by triangles), $[-10^\circ, 10^\circ]$ (shown by crosses) and $[20^\circ, 40^\circ]$ (shown by circles)	82
4.3	Visualisation of the hierarchical pose model.	83
4.4	Means corresponding to each of the PPCA models defined by the hierarchical pose model with their respective yaw labels.	84
4.5	Visualisation of linear relationship for the pitch estimate for an example cluster using discretely sampled training data at 10° intervals.	86
4.6	Visualisation of the linear relationship for the yaw estimate using discretely sampled training data at 20° intervals.	87
4.7	Selected training samples from the pose estimator dataset.	87
4.8	Visualisation of the eye and mouth centroids with the latter calculated by intersecting two lines formed by the landmarks on lip and jaw outline.	89
4.9	Denser shape model consisting of 74 landmarks (top), sparse set of 14 landmarks (bottom) with corresponding estimated pose angles (yaw,pitch,roll).	90

4.10	Pose estimates obtained from the original pose model (top) and from the yaw constrained one (bottom).	91
4.11	Section of pitch rotation plot: original jagged curves (green solid line) and their smoothed out counterparts (red dashed line).	92
4.12	Results of continuous pose estimation experiment.	93
4.13	Selected samples from the original sequence (left image) with corresponding frames from the synthesised avatar (right image).	94
4.14	Visualisation of triangulation when warping using a linear model: (a) base (mean) triangulation, (b, c) extreme view triangulation.	96
4.15	Visualisation of triangulation when warping using a hierarchical model: (a) base (mean) triangulation, (b, c) extreme view triangulation.	96
4.16	Shape and expression-free appearance with full PDM mask (top) and shape-free with pose PDM mask (bottom).	97
4.17	Examples of morphing (synthesising) facial expressions into extreme virtual views. The top rows in (a), (b), (c) were computed using the HLVM model. The bottom rows were computed using the PCA model with visible kinks at extreme 3D views (profile views) due to the non-linearities present.	98
4.18	Visualisation of the cluster membership for the shape bases used in the synthesis (top and middle rows) together with the synthesised samples (bottom row).	99
4.19	Distortions due to the pose changes and self-occlusion. Top row: original images, bottom row: frontal view warped images.	100
4.20	The mirroring process. Original images (top row) and resulting pose corrected frontal view warped images (bottom row).	100
4.21	Different weight vector representations for different yaw rotation values.	101
4.22	Selected frames from the experiment in fitting AAM onto the extreme pose view. Top row corresponds to the original AAM model formulation and bottom row to the pose corrected model.	102

5.1	General overview of the process in our system.	107
5.2	Selected frames from two expressions demonstrating gradual change for grin (top row) and fear (bottom row) and the severities associated with them using our method (Severity Criterion (SC)) and naive approach based on (Mahalanobis Distance (MD)).	111
5.3	An example of morph bases (left to right) for neutral, smile, grin, surprise/fear, anger for two different avatars (top and bottom).	112
5.4	Label assignment together with the corresponding ground truth and selected keyframes for test sequence T1 where the top row corresponds to rule-based, middle to holistic and bottom to BN approaches.	114
5.5	Label assignment together with the corresponding ground truth and selected keyframes for test sequence T2 where the top row corresponds to rule-based, middle to holistic and bottom to BN approaches.	115
5.6	Final label assignment scores for test sequences T1 and T2	116
5.7	Severity curves for test sequence T1 (bottom row) together with corresponding label assignment scores for rule-based approach (top row).	118
5.8	Severity curves for test sequence T2 (bottom row) together with corresponding label assignment scores for rule-based approach (top row).	119
5.9	Selected frames from experiments on expression classification and avatar animation with corresponding labels and severity (as a percentage). Each of the images shows the tracked frame with AAM mask superimposed on it (left), and corresponding synthesised avatar (right).	120

List of Tables

3.1	Proportions of intrinsic eye states present in the training set.	69
3.2	Confusion matrices of the left eye model (a) and the right eye model (b).	71
3.3	Proportions of intrinsic eye states for T1 and T2 test sequences.	72
3.4	Confusion matrix of the overall eye state classification for test sequence T1 (a) and T2 (b) for the combined eye model.	73
3.5	Proportions of intrinsic mouth states present in the training set.	75
3.6	Proportions of intrinsic mouth states for T1 and T2 test sequences.	75
3.7	Confusion matrix of the mouth state classification for T1 (a) and T2 test sequences in a model built using a person-specific dataset	79
5.1	Table comparison of label assignments methods.	116

Acronyms

AAM Active Appearance Model.

AFEA Automatic Facial Expression Analysis.

ASM Active Shape Model.

AU Action Unit.

BN Bayesian Network.

CDF Cumulative Distribution Function.

EMFACS Emotion Facial Action Coding System.

FACS Facial Action Coding System.

FACSAID Facial Action Coding System Affect Interpretation Database.

FAP Facial Animation Parameter.

GMM Gaussian Mixture Model.

GPA Generalised Procrustes Analysis.

HCI Human-Computer Interaction.

HLVM Hierarchical Latent Variable Model.

HMM Hidden Markov Model.

HPCA Hierarchical Principal Component Analysis.

IAM Independent Appearance Model.

ICA Independent Component Analysis.

IR Infra-Red.

KPCA Kernel Principal Component Analysis.

LBP Local Binary Patterns.

LDA Local Discriminant Analysis.

LFA Local Feature Analysis.

LLE Locally Linear Embedding.

MD Mahalanobis Distance.

MU Motion Unit.

N-SVD Normalised Singular Value Decomposition.

NLPCA Non-linear Principal Component Analysis.

NN Neural Network.

PCA Principal Component Analysis.

PDF Probability Distribution Function.

PDM Point Distribution Model.

PPCA Probabilistic Principal Component Analysis.

RMSE Root Mean Square Error.

SC Severity Criterion.

SVD Singular Value Decomposition.

SVM Support Vector Machine.

VSR Valid Shape Region.

Chapter 1

The Introduction

Facial expressions provide one of the most powerful and fastest means of relaying emotions between humans. As faces are the most easily recognizable feature of any individual, recognition by facial identification is faster in most people than for example by name. Neuroscientists suggest that there is presence of a region in the human brain specifically dedicated to face recognition, and that this region is further subdivided into a task oriented components responsible for emotion, identity and gender recognition [48]. Our brain is a beautiful and extremely powerful thing and it is not surprising that we take for granted the ability to process efficiently and seamlessly all of this information in real time.

Rapid development of hardware in the last two decades has resulted in an ever greater increase in processing power, the emergence of the Internet as a global communication medium and the ever growing popularity of personal computers and portable devices have opened up new and interesting areas for vision related research. Although we are far away from matching the ability and performance of our brains, the ability to re-create facial expressions using synthetic counter-parts, allowing the simulation of direct visual communication between humans using computer devices, and the in depth understanding of human nature enables the normally emotionless machines to exhibit, to some

degree, personality and emotions that increasingly seem more real than ever.

Because of this we can see a change in the way that we perceive Human-Computer Interaction (HCI) designs, with traditional computer-centered setups involving keyboard and mouse, are replaced by a notion of “ubiquitous computing” where human-centered designs are at the forefront [124]. Due to such an increase in interest, and the amount of research carried out, new opportunities and application fields have emerged where there is a need to understand or quantify facial expressions is much desired. Some of the potential fields include:

- **Computer Animation:** From the humble beginnings of *Tron*, through to the pioneering work of *Pixar* in this field, and because of its visual nature and its global exposure, this area has made the most noticeable progress, and features in almost every aspect of the entertainment field today.
- **Computer Games:** Due to the establishment of the internet as a global communication medium this field has evolved from that of a lonely, or personal, experience to a social phenomena, where a multitude of players from all around the world meet and immerse themselves in fantasy realms. With the emergence of Massively Multiplayer Online Role-Playing Games (MMORPG) that facilitate and encourage communication between players the need to visually convey, or determine their emotions as the players react to unfolding events has never been more desirable.
- **Virtual Meeting/Chat Rooms:** From the dawn of time we have relied on person to person communication and social interaction. Text based communication has been replaced by visual interaction, where in Instant Messaging Clients (IMC) and specifically designed social virtual worlds, such as *Second Life*¹, the addition of facial expressions would complete the experience.
- **Monitoring:** Ability to determine certain behaviours and their extent can be useful, and can provide insight, or cues on further actions. These states might include pain [69], fatigue [55] or deceit [130].

¹Second Life. <http://secondlife.com>



Figure 1.1: Gollum character in Lord of the Rings.

- Education: Automated systems that can recognise the underlying emotional states of tutees, such as interest [119], and the opportunity this insight provides for the system to adapt or tune its programming according to those recognised states.

However, in our great quest for realism, by large the best results are still obtained by employing expensive or complex motion capture systems, as the ones used on the set of “Beowulf”, or relying on skills and experience of animators to transfer facial expression dynamics into synthetic counterparts, where the prime example is the animation of Gollum’s face in “Lord of the Rings” (Figure 1.1).

Recent interest and advances in computer vision research have produced works that focus on automated and non-intrusive approaches to extract and define relevant information that does not depend on expensive and time consuming setups. However many difficulties of the real world scenarios are still present. They include:

Tracking problems These will occur due to the ubiquitous noise, occlusion, shadows and low quality of the input source. Their presence might cause loss of tracking or inappropriate information being extracted hence causing erroneous classification and animation. In this thesis we do not directly focus on tracking phase although it is particularly important as it provides the

foundation which our approach is built on.

Complexity of expressions Expressions are something more than just a contraction of facial muscles, they can convey the current emotional state of the person and can open a new dimension in human-computer interaction. They are also extremely diverse, as no two people exhibit the same expression in the same way. This makes it very difficult, as trade-offs have to be made, regarding generality, complexity and performance of the models we need to use.

Realistic animation Because people learn how to recognise and interpret faces and facial expressions from the day they are born, even the slightest imperfections can be easily spotted. This makes our quest for realism very challenging [76].

1.1 Approach

The goal of this thesis is to address the problem of extracting the dynamics of facial expressions and providing a parametric description that can be subsequently used to animate a synthetic counterpart. We can think of it as a process of transition between the real human face and the resulting synthetic counterpart. In this thesis we study the following problems:

1.1.1 Feature Representation

A very important step for any successful vision system is the choice of underlying feature representation. Due to the complexity of the facial expressions and the multitude of available stimuli the choice is difficult and crucial as it defines the building foundation. The thesis work begins by introducing the Active Appearance Model (AAM) which consists of a holistic shape and appearance combined elegantly using a PCA statistical model. We focus on their person-specific variant, which provides better performance [46], and a more optimal basis from which to capture subtle dynamics of facial expressions under a sparse training set.

Most of the existing approaches adopt a holistic representation, where the non-linearities caused by the motion of facial features and their intra-feature correlations have to be accounted for in the

resulting expression space. Also such a representation can encode redundant information that is unnecessary, or even detrimental to the underlying task. We introduce a hierarchical decomposition where the face is separated into eyes and mouth components which are the most salient regions on the human face. This approach allows us to represent each of these regions according to their intrinsic functionalities rather than the final expression labels. We focus specifically on the shape component of the AAM, referred to as the Point Distribution Model (PDM), because of its resilience to the illumination changes and underlying low dimensionality.

By taking advantage of facial symmetry we utilise a single eye model to represent the variation of both of the eyes. The underlying subspaces are modelled using the Hierarchical Latent Variable Model (HLVM) [13]. Whilst retaining PCA properties, it defines probability density model, that allows the discovery of internal within cluster structures, but most importantly it allows modelling of non-linear manifolds with a combination of localised submodels. We demonstrate the advantages of such a model over the PCA approach.

1.1.2 Pose Estimation

Head motion is an inherent component of facial dynamics. The continuous head movement, every tilt, shake or nod is what we learned to perceive, and are used to, as a constant companion of facial expressions. Ability to capture and model such movements will compliment, and enrich the resulting information. A prime example of its importance is *Pixar's* short feature film *Luxor Jr.*² where a common household object was given human-like characteristics through the inclusion of such subtle movements.

Due to the highly non-linear variations, linear mapping is not sufficient to represent pose changes. We introduce a pose model, which is based on the first level of the HLVM hierarchy, which can model the non-linear space with a combination of linear components. By just using the shape component and appropriate underlying features, the influence of facial expressions on the resulting pose estimation is minimised, something that would be difficult if the appearance component was also

¹Luxor Jr. <http://www.pixar.com/shorts/ljr/index.html>

considered. Based on a sparse and discretely sampled training set our approach is able to estimate pose through its probabilistic framework. We exploit our previously introduced hierarchical decomposition, and show that through sub-sampling of the denser shape model we can derive continuous pose estimates. As we do not utilise any dynamic information this method is fast and can deal with large jumps and discontinuities.

We explore the usefulness of our pose model to the synthesis of facial expressions at arbitrary viewpoints under an independence assumption of shape and texture. For a linear shape model warp basis, deformations at extreme viewpoints will result in texture artifact and distortions. Our model is able to overcome these problems. The choice of features, driven by minimising the effects of expressions, has a reverse effect where most of the variations are kept in the appearance space and the shape only serves as a warp vessel. Finally we investigate the effect of pose information on fitting of AAM where at extreme views self-occlusion causes some of the information to be unavailable. We employ a dynamically generated, and pose-dependent weight vector that constraints the calculation of appearance difference at these views and improves the performance.

1.1.3 Dynamics Extraction and Parameterisation

Most of the existing approaches in Automatic Facial Expression Analysis (AFEA) focus on only determining the underlying facial states, ignoring the underlying gradual changes and treating it as an on/off process [41]. Given our previously defined hierarchical models we explore the concept of fusion to combine the intrinsic functionality of eye and mouth components into final expression labels. We investigate the use of rule-based classifiers and compare them with a Bayesian Network (BN) similar to Cohen et al. [20], but with the main focus on the independence between regions rather than features, and holistic representations of Huang and Huang [53] and Liu et al. [71]. Given the extracted final expression label, we calculate the corresponding intensity as a gradual change of the combination of intrinsic intensities of the hierarchical subcomponents. These are extracted using existing probability density models.

By using the AAM tracker on sequences containing a mixture of facial expressions interleaved

by speech fragments we show that our method offers better performance over the BN and holistic approach. Finally we combine extracted expression labels together with their intensity and pose information and apply its parametric form, in a morph-based fashion, to a synthetic head avatar.

1.2 Contributions

The main contributions are:

- A hierarchical face representation has been introduced that utilises redundancy implied by the symmetry of the human face. This representation reduces overall complexity of the model. A statistical model that is built upon this hierarchical representation allows us to accurately recognise expressions and extract their rate of change [123, 122].
- A 2D pose model built using a sparse set of training samples provides invariance under facial expressions, generalises to continuous and unseen samples. The model serves as a basis for synthesis of facial expressions across arbitrary views and can improve the AAM fitting process [121, 120].
- A fusion framework that draws from our hierarchical models to produce expression labels, aids in the definition of the resulting expression intensity and applies resulting parameterisation to animate a synthetic head avatar. [123, 122, 120].

1.3 Overview of the Thesis

Chapter 2 Review of previous work on facial expression analysis.

Chapter 3 Introduction of hierarchical facial decomposition into sub-components and their representation using the Hierarchical Latent Variable Model (HLVM).

Chapter 4 Investigation into pose estimation based on our previously introduced hierarchical representation and its usefulness in synthesis and tracking.

Chapter 5 Overview of the framework for fusing hierarchical subcomponent information with rule-based classifier and resulting expression intensity estimation. Application of parameterised data to synthetic head counterpart.

Chapter 6 Conclusions and future work.

Chapter 2

Literature Review

Early work by Suwa et al. [101] identified many problems associated with facial tracking and expression recognition. Recent advances in hardware development and huge increases in the processing power of computers have triggered considerable interest in the research community. However development of vision systems that automatically understand and synthesise facial expressions is still a rather difficult and daunting task. A huge body of work is already present in the areas of Automatic Facial Expression Analysis (AFEA) [41, 87, 88, 106, 72, 85, 124, 99]. The primary focus of these studies is aimed at providing relevant labels, the underlying context of expression dynamics however is mostly overlooked. We can divide the process of expression dynamics analysis into the following stages:

- Face acquisition - This step is necessary to find or detect a face in an input source, which can be an image or a sequence of images. The problem of *face detection* has been largely addressed in the literature [50, 131, 117]. Once the face has been detected, usually some alignment is performed. Also pose estimation can be performed if significant head motion is present.
- Feature extraction - Once the face has been located, this step will extract and formulate the features necessary to represent facial expressions. The exact representation of features can

vary. It will form the basis for stages that follow.

- Semantics (or expression dynamics) formulation - will extract and formulate the information needed to describe *facial state* in a parametric form, which can then be used for avatar animation. Facial expressions underneath are more than just descriptive labels that we usually associate with them. They are a direct manifestation of our emotional state, and are affected by the surroundings and the context they were used in [72]. On the surface they are an intricate interplay of facial features driven by muscle contractions. The way and the rate they evolve over time, including rigid head motion associated with them, are the integral part of their dynamics. Since no two people exhibit them in the same way, and due to our susceptibility to even smallest imperfections, the ability to capture and learn those underlying semantics accurately poses a great challenge.

In the following section, we review the existing work on feature representation (Section 2.1) and semantics formulation (Section 2.2).

2.1 Feature Representation

In order to successfully recognise expressions we need to choose the appropriate representation of “features”. We refer to features as attributes used to describe *the model* of the face, rather than a specific representation thereof, such as fiducial points of prominent landmarks on the face like nose, mouth or eyes. The choice of the appropriate features is important for reasons such as computational efficiency, discriminative power and resilience to missing and incomplete data [93]. Most importantly such choice will have a direct effect on the methods used to extract information from the input source and the approaches used to process that information. We can subdivide the representation into two different levels: conceptual and data-based.

2.1.1 Conceptual Level Representation

On the conceptual level, features can be defined as holistic (i.e. as a whole unit) or localised, to reflect changes in specific regions of the face. Holistic approaches treat the face as a whole unit, usually by combining all of the information such as shape, appearance and motion into a single monolithic unit, and process them as such. The underlying feature vectors usually have high dimensions, especially when dealing with the appearance data. They embed information, which might be irrelevant for expression recognition task, and could have a direct impact on the classifiers ability to uncover its true nature [84]. More importantly they implicitly encode correlations between facial features, which can cause non-linearities in the resulting expression space. These correlations are important [21, 100] as they determine the underlying facial states, but do not have to be accounted for at this particular stage. Similarly, they will also implicitly encode global information of the underlying dynamics.

Localised features are associated with specific parts of the face. These parts are usually based on the areas that are mostly prone to change, or contain the most relevant information for facial expression recognition. Psychophysical experiments conducted by Cunningham et al. [27] suggest that changes only in certain regions of the face, are sufficient to successfully recognise facial expressions. In their work, they focused on the eye and mouth regions, which are the most intuitive and salient regions of the human face. They concluded, that these regions, either on their own or in combination can be used to recognise facial expressions. Furthermore these regions have been widely used and relied upon in computer facial animation [42, 81, 63]. Nusseck et al. [81] examined the concept of necessity with respect to different facial regions towards successful recognition of expressions. Their findings suggest that for some of the expression categories individual regions of a face alone can be used to successfully represent the expressions. Berisha et al. [10] have also highlighted the importance of those regions and their influence on the global perception of the face. Rather than performing the analysis from the perspective of recognition, or functional importance of the regions, the significance of these regions was described in the context of sources of variation in the facial image.

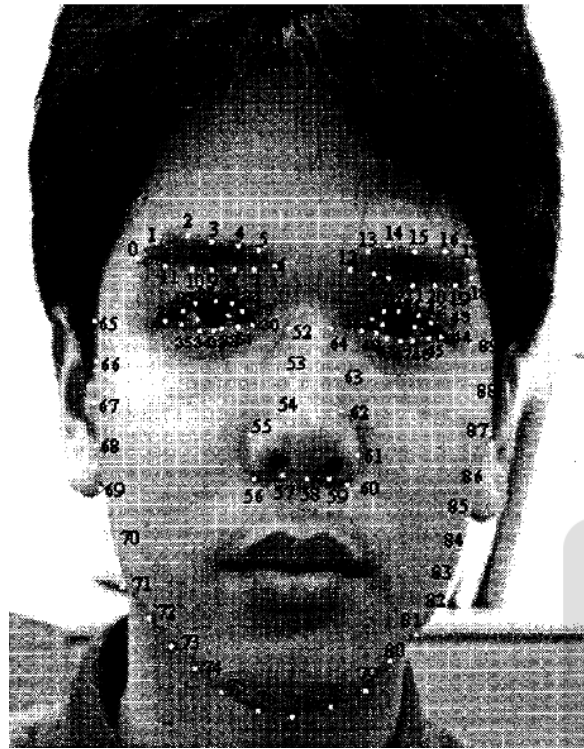


Figure 2.1: The 90 feature points used for PDM representation. From Huang and Huang [53].

2.1.2 Data Level Representation

On a different level, data-representation features can be divided into two main categories: geometric and appearance-based [106]. Geometric features usually define the location and/or shape of facial components, such as the mouth, nose or eyes. Appearance-based features represent the textural information of the face, such as hair or skin, and its visual characteristics such as furrows or wrinkles. This level is orthogonal to the previously defined conceptual level such that any data-level representation can be either holistic or localised.

Geometric Features

Huang and Huang [53] used 90 landmark points placed on the face (Figure 2.1) combined with parabolic curves for upper and lower lip shapes. Resulting shape was represented by a Point Distribution Model (PDM) to capture the underlying holistic shape variations using Principal Component

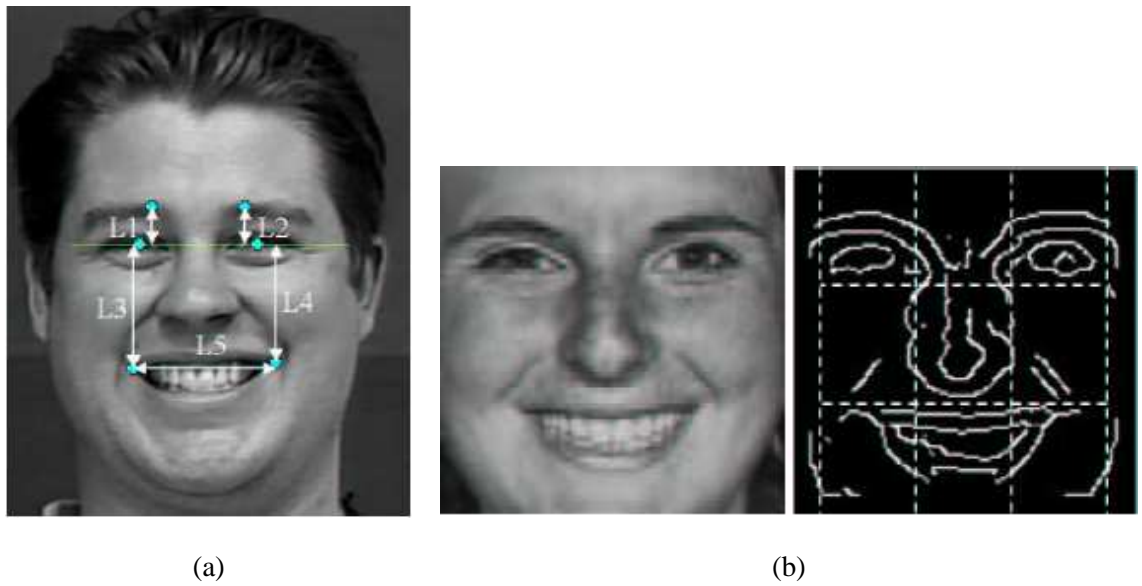


Figure 2.2: (a) Location features, where $L1, L2, L3, L4, L5$ correspond to distances between six facial features. (b) Shape features with normalized image on the left and zones of the edge map on the right. From Tian et al. [105].

Analysis (PCA) under a linear relationship assumption, which tends to fail in the presence of non-linear variations such as those caused by large facial or pose variations. No rigid head motion, or illumination variations should be present. Similar shape model based approach was used by Chang et al. [16, 17], but with 58 facial landmarks. Tian et al. [105] used a combination of six location and shape features from frontal, or near-frontal views. The six location features consisted of the eye corners, eyebrow inner end-points, and mouth corners, and were represented by five parameters defined by distances between the features in question. Shape features were used for the mouth area and defined by applying an edge detector to a normalised face to produce a 3×3 edge map. Then the mouth shape features were computed from zonal shape histograms of the edges of the mouth region. Figure 2.2 shows the location features (a), and shape features with the normalized face on the left and corresponding edge zone map on the right (b). Valstar and Pantic [110] used a set of 20 fiducial facial points (Figure 2.3(a)) plus the location of both irises and the centre of the mouth in a frontal view which locations were detected using Gabor-feature based boosted classifiers. Terzopoulos and

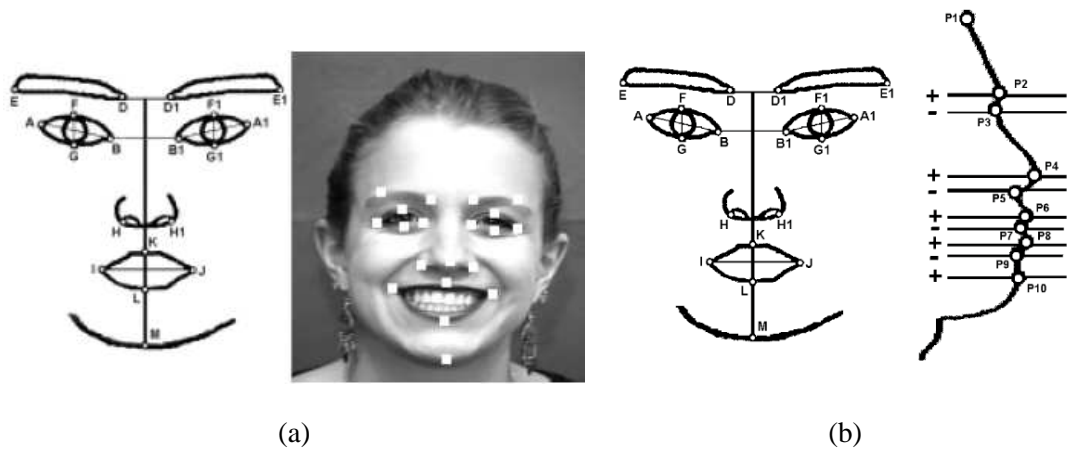


Figure 2.3: (a) 20 frontal fiducial facial points (from Valstar and Pantic [110]) (b) 19 frontal and 10 profile fiducial facial points (from Pantic and Rothkrantz [89]).

Waters [103] used 11 active snakes (also known as deformable contours) to represent lips and facial features. Their approach required the subjects to wear feature enhancing make-up therefore making it unsuited for real world environments.

Although the majority of existing approaches deal with frontal views only, multiple, or alternative views can be employed in order to bootstrap and help disambiguate issues related to the feature extraction process. Pantic and Rothkrantz [88] used frontal and profile views with 30 and 10 location points respectively. Multiple detectors were used for each of the prominent facial features, and the best representation was chosen based on knowledge of the facial anatomy and the confidence of each of the detectors. This provided some measure of redundancy, but was limited solely to the detection stage and required manual initialisation. Pantic and Rothkrantz [89] utilised frontal and profile views with 19 and 10 points respectively (Figure 2.3(b)). However the practicality of their approach was impaired by the need to wear head mounted camera rig. In their following work Pantic and Patras [86] focused on a profile view only using 15 points.

Besides the location and/or shape of the features, cues such as motion or displacement can be added. These can only be used with images sequences where such information is available. Kimura and

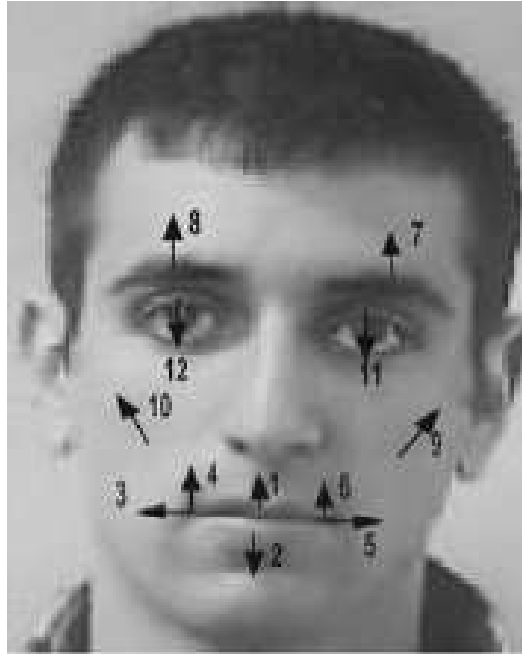


Figure 2.4: Motion Units (MUs) representation. From Cohen et al. [21].

Yachida [58] used an elastic net model, or more precisely, the motion vectors caused by its deformations. Otsuka and Ohya [83] used optical flow to estimate the motion of localised face regions such as mouth and right eye. The exclusion of the left eye region meant that the changes that occur in single eye only cannot be adequately represented. Also the facial symmetry assumption regarding the range of motion in both parts of the face cannot always be guaranteed. Cohen et al. [21] used an explicit 3D wireframe model of the face to represent direction and magnitude of the displacement of 12 facial features. He called them Motion Units (MUs) (Figure 2.4). Essa and Pentland [40] used optical flow to estimate facial movements. Those movements were then constrained and refined recursively by a detailed physical face model of Platt and Badler [92]. Lien et al. [67] used motion vectors extracted from coarse-to-fine pyramidal optical flow feature tracking, dense optical flow for holistic face motion, and spatio-temporal domain gradient information for modelling of furrows in the skin. Lee and Xu [65] also applied pyramidal optical flow to calculate the displacement of 19 feature points placed around facial landmarks such as the eyes, brows, nose and mouth. In general, flow-based techniques

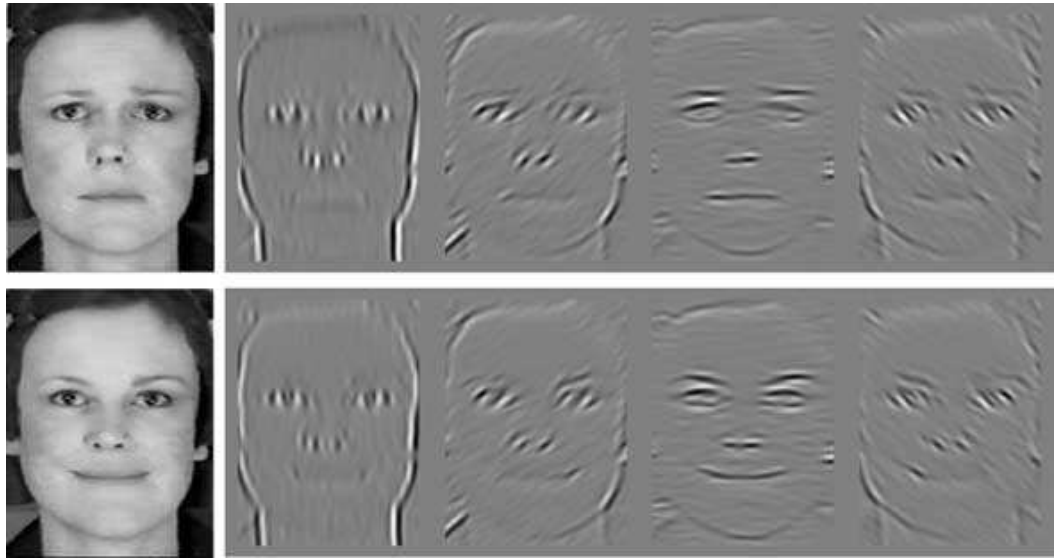


Figure 2.5: Facial features resulting from Gabor wavelet convolution. On the left are two facial expressions, and on the right Gabor representations representing four differently oriented kernels. From Fasel and Luetttin [41].

are computationally intensive and easily influenced by variations in lighting, non-rigid motion and are sensitive to image registration and motion discontinuities [128].

Overall, geometric features provide low dimensionality data and a conceptually easy way to model and describe expressions. They are tolerant to variations in illumination, and to some extent also to inter-person differences. Unfortunately they can't easily encode subtle skin changes such as wrinkles or furrows caused by facial expressions. Their main disadvantage is that they rely on accurate and reliable detection and tracking, which in many situations cannot be guaranteed. They do not deal very well with occlusion, and large pose variations. Also in low resolution images and real-world environments such information might not be easily available [105, 106].

Appearance-based Features

Gabor filter based features [70, 5, 108, 110, 68, 6, 2, 71] are the most widely used appearance feature representations due to their resilience to variations in lighting and to small shifts and deformations [85]. Figure 2.5 shows sample features obtained from Gabor wavelet convolution. They can be applied in holistic fashion, or to specific locations, whereby Gabor filters are placed at selected

locations on the face to limit computational complexity [112].

Statistical analysis, such as Principal Component Analysis (PCA) [109], Local Feature Analysis (LFA) [28], Local Discriminant Analysis (LDA), Independent Component Analysis (ICA) [28] have also been widely used. They were pioneered by the “eigenfaces” work of Turk and Pentland [109]. Padgett and Cottrell [84] explored the concept of local principal components, or eigenfeatures, in which windows were placed around facial feature regions (such as eyes and a mouth). Eigenfeature based representation produced better results than its holistic counterpart, but required normalised image data and was susceptible to differences in the mouth and eye structure. Donato et al. [28] compared the use of several techniques such as holistic spatial analysis, including PCA, ICA, LFA, LDA and approaches based on the outputs of local filters, such as Gabor wavelet representation and local principal components. The best results in recognising facial expressions were obtained using Gabor wavelet representations and ICA. This experiment demonstrated the superiority of localised representation.

Since appearance feature numbers and their dimensionality can be large, this implies high computation costs, for example in Gabor convolution using a large number of features. To overcome this problem, various feature selection techniques have been proposed to select a subset of the most effective ones. Littlewort et al. [68] and Bartlett et al. [6] employed AdaBoost to select the best subset of Gabor filters. Valstar and Pantic [110] compared GentleBoost feature selection against AdaBoost and by using their dataset the former outperforms the latter. Local Binary Patterns (LBP) provide an alternative to the Gabor based representation. First introduced by Ahonen et al. [1] and subsequently adopted by Shan et al. [98], they offer lower computational cost and tolerance against illumination variation, whilst retaining sufficient internal feature information. Shan [100] compared AdaBoost with Conditional Mutual Information (CMI) based Boosting using LBP features, noting that AdaBoost produced better results.

The general criticism of appearance-based features is that they are more susceptible to illumination variations and inter-person differences, unless a substantial quantity of features and large varied training datasets are used [85]. Zhang et al. [129] have shown that Gabor-filter based appearance

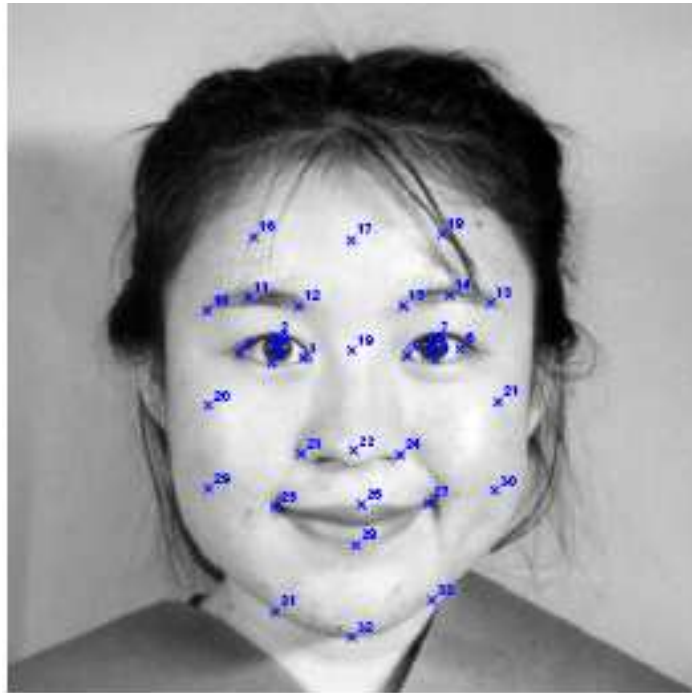


Figure 2.6: 34 fiducial points representing facial geometry. From Zhang et al. [129].

features perform better than geometric features alone. This is understandable as they can generally encode more information compared to their geometric counterparts. However, study conducted by Pantic and Bartlett [85] shows that this is not always the case. Work of Pantic and Patras [86] who used a geometric feature representation performs similar or better than some appearance methods, notably that of Bartlett et al. [7].

The obvious way forward is to combine both geometric *and* appearance features together, hence utilising both their strengths and minimising their individual weaknesses [124]. Zhang et al. [129] used 34 fiducial points selected manually on the face (Figure 2.6) and applied a set of multi-scale and orientation Gabor filters at those points. Tian et al. [104] used a multi-state face component model: a three-state lip model (with open, closed and tightly closed states), a two-state eye model (with open or closed states), and one state model for the brow and cheek. In addition, appearance features from the eye and nose area were incorporated into a two state model (present or absent). Zhang and Ji

[126, 128] used a set of 26 facial features around the eyes, nose and mouth and a set of transient features similar to that of Tian et al. [104].

The Active Appearance Model (AAM) [33, 77, 114, 73, 59, 11] is another example of a model relying on both geometric and appearance-based features. AAM uses a holistic representation with combined shape and appearance features modelled in a PCA space. Similarly to PDM, which is a component of AAM, it models both shape and texture under a linear relationship assumption, and in the same way it tends to fail in the presence of non-linear variations caused by large facial or pose variations [43]. Matthews and Baker [77] improved the performance of AAM in terms of speed and accuracy by using an inverse compositional fitting algorithm. However their method can only work with an independent AAM, in which shape and appearance are not combined, and increases the dimensionality of the resulting model. Xiao et al. [114] utilised both 3D as well as 2D information in what they call a combined 2D+3D AAM. This method uses 3D information to bootstrap the fitting process and is more resilient to occlusions, but it requires an additional 3D input training data. Unfortunately the AAM approaches require very large and exhaustive training sets in order to accurately capture and model inter-person and intra-person expression-specific variations. Gross et al. [46] introduced the concept of Person-specific AAMs. This improves the performance and reduces the number of training samples by removing inter-person variation, but requires training separate models for each individual person. Zhang and Cohen [125] investigated the lack of flexibility in the holistic representation of AAM and presented a component-based approach, where separate models for eye and mouth-nose regions were combined with a global representation providing more flexibility, accuracy in shape localisation.

2.1.3 Expression Manifold Representation

The dimensionality of a face input space is defined by the features used to represent it. It will range from tens in the case of geometrical features to thousands for appearance features. Given high dimensionality of the face input space, face representations lie intrinsically on lower dimensional manifold [102, 97]. An expression can be represented as a point on such a manifold and variations, or changes

in those expressions can be represented by a path in such non-linear manifold, which is of intrinsically lower dimensionality than its corresponding input space [16, 17, 97].

Identification of a global, concise and accurate representation for all possible facial expressions, is crucial for modelling facial dynamics. Subspace analysis methods such as PCA and LDA have been used to model manifolds, but they fail to find the true structure of the data due to its underlying non-linearity [102]. In some cases PCA can be sufficient when a small set of significantly different expressions is used, hence keeping resulting non-linearities to the minimum. For example Kimura and Yachida [58] represented anger, surprise and happiness in such way. Liu et al. [71] and [53] employed a PCA based representation combined with a Gaussian Mixture Model (GMM) to represent the distribution of facial expressions. While the addition of a GMM produces better results compared to PCA alone there was significant overlap of clusters in the manifold. This is understandable, as GMM can only provide constraints on the shape variations to aid the classification, but it cannot reliably model the underlying non-linear data variation, as its built on top of PCA representation and inherits its underlying flaws. Various extensions to PCA have been proposed to allow representation of non-linear manifolds, such as Non-linear Principal Component Analysis (NLPCA) or Kernel Principal Component Analysis (KPCA) [43]. Although they overcome the linearity problem by utilising higher dimensional subspace, these methods capture the overall variance of the data which might be the most optimum solution. Moghaddam [78] investigated use of linear and non-linear techniques such as PCA, ICA, NLPCA and KPCA for manifold modelling in the context of face recognition. Their results show that in terms of matching accuracy, KPCA outperforms PCA by margin of 10% but it is surpassed by Bayesian methods such as Probabilistic Principal Component Analysis (PPCA). Heap and Hogg [49] introduced the concept of Hierarchical Principal Component Analysis (HPCA) where combination of local linear sub-models were used to represent non-linear variations in PDM. Although the work was in context of shape modelling and underlying constraints, the principals are applicable to generic scenarios. Their model consisted of a global PCA model in the root of the hierarchy, and linear combination of local PCA models constructed in the global, or root, PCA subspace. Tenenbaum et al. [102] used Isometric feature mapping (Isomap) to discover meaningful

structure from high dimensionality data, while Chuang et al. [19] used shape and texture statistical representation similar to Cootes et al. [25] and a bilinear model to represent the facial expression space. Symmetric and asymmetric formulations were employed to model speech and three facial expressions. Du et al. [31] presented a method of mapping expressions to a 2D Valence-Arousal emotion space, based on the 3D Valence-Arousal-Control model of Russell and Mehrabian [95] using linear mapping. Given the non-linear nature of the data, such a mapping is not sufficient to efficiently model the manifold. Chang et al. [15] compared two types of embedding for expression manifold modelling: Locally Linear Embedding (LLE) and Lipschitz embedding. Chang et al. [16] also presented a probabilistic framework for recognising and synthesising six basic expressions using enhanced Lipschitz embedding. In their following work Chang et al. [17] proposed an Active Shape Model (ASM) representation and existing probabilistic framework to model six basic expressions. As they modelled a global, generic expression manifold, their training set only included two subjects, which barely provided sufficient validation. Shan [100] used Locality Preserving Projections to model a generic manifold in the LBP appearance space to account for more detailed feature variations. They point out that sparse geometrical representation is not sufficient as it does not capture detailed features such as wrinkles or furrows, and that having a separate manifold for each subject does not provide enough generalisation.

2.2 Semantics Formulation

In order to successfully classify an expression one needs to define a “dictionary” that describes our facial expression state. It is worth pointing out that facial expression analysis is not analogous to emotion analysis [106, 72]. In order to understand emotions, higher level knowledge is required and additional factors such as context, body language, cultural elements and sound should be considered. There are two distinct approaches: sign and message judgment [22]. Subsets of those are sometimes referred to as emotion (affect) based, and muscle (Action Unit) based [106, 100, 85]:

- * *Message judgment* - message judgment tries to describe facial expressions in terms of inferred emotions. During an extensive study Ekman [35] proposed six categories of emotional ex-



Figure 2.7: Universal expressions (from left to right): Fear, Joy, Disgust, Surprise, Sadness, Anger. From Kanade et al. [56].

pressions, referred to as the basic, or universal expressions: happiness, sadness, surprise, fear, anger and disgust (Figure 2.7 shows examples of each). His experiments suggest that these basic expressions are universally displayed and recognised across different cultures. This is by far the most popular method and has been widely adopted in the field of automatic expression recognition. There have also been attempts to recognise other emotional states such as interest [119], pain [69], fatigue [55], deceit [130], and conversational components such as agreement, disagreement, thinking, cluelessness and confusion [81].

* *Sign judgment* - Sign judgment attempts to describe facial expressions in terms of surface behaviour such as facial component movement or change. The Facial Action Coding System (FACS) [36] is considered to be the most popular for analysing facial activity. It defines the expressions in terms of 44 atomic Action Units (AUs) in which 30 of those correspond to movements of particular muscle groups and are partitioned to two sets, corresponding to upper face and lower face (12 for the upper face and 18 for the lower face). The remaining 14 AUs have anatomically undefined basis and are referred to as miscellaneous actions [56]. Using those rules a particular expression can be decomposed into a single AU or the combination that describes that particular expression. More than 7000 combinations have been observed [57]. Figure 2.8 illustrates some of the AUs for the upper part of the face. However FACS in itself does not define a way to translate the AUs into labelled expressions. Extensions such as the Facial Action Coding System Affect Interpretation Database (FACSAID) [38] or Emotion Facial Action Coding System (EMFACS) [37] allow translation of AUs into predefined
















<i>NEUTRAL</i>	AU 1	AU 2	AU 4	AU 5
				
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
				
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
				
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 2.8: Some Action Units corresponding to the upper part of the face. From Tian et al. [104].

emotional categories. Alternative system is the Facial Animation Parameter (FAP) [82] coding system developed for the MPEG-4 standard. It defines a set of 84 feature points (FPs) placed at predefined locations on the face. Facial movements are represented by FAPs which are defined as FP displacements with respect to a neutral state, and are measured in FAP units which correspond to the distances between key features. Although this system does not define the way by which we formulate expressions, it does define the template on which the data should be represented. It is based purely on facial feature points, unlike FACS, which is driven by facial muscle movements.

2.2.1 Pose Information

Changes in pose are common by-products of facial expressions. Some of the expressions, such as agreement or disagreement, can be represented by head movements alone [27]. Since our head rarely

remains still, and its motion directly influences expression dynamics, it is imperative to extract this information in order to provide more complete and accurate semantical description. Relatively few works in the literature try to model pose together with facial expressions, or take such information into consideration in the semantics formulation process [124]. The rigid head motion is intrinsically embedded with non-rigid local feature motion and the main challenge is in separation of the two. Usually the rigid motion is estimated first, and the non-rigid second, entailing that the face is a rigid object without taking into consideration expression changes [132]. Bascle and Blake [8] investigated coupling between pose and facial expressions. For the expressions, they used a linear model built from key expressions, and the pose estimation method used affine projection with parallax. Combined influence of both was represented with a bilinear model and de-coupling was achieved using Singular Value Decomposition (SVD). The main focus of their work was on the removal of pose influence in the expression recognition process, and pose information was not explicitly defined. Also the subject were required to wear enhancing make-up. Zhu and Ji [132] proposed an improved approach based on a Normalised Singular Value Decomposition (N-SVD) to recover head pose information in an analytic fashion. Constraints were added using non-linear techniques imposing orthonormality condition on the pose parameters. The evaluation on a synthetic data, where the original feature points were displaced by applying different levels of Gaussian noise, shows significant improvement over Bascle and Blake [8]. Sarris et al. [96] used a 3D model and optical flow to estimate pose information from video sequences. The assumption was that the 3D model could be adequately fitted in the very first frame. Eisert and Girod [34] also employed 3D models. Both methods fitted into the FAP framework. Gu and Ji [47] used Infra-Red (IR) information to define several properties such as inter-pupil distance, size or orientation, which were subsequently modelled using PCA to form a Pupil Feature Space (PFS). Head pose was determined by mapping those parameters into PFS. A similar technique was used by Zhang and Ji [127]. Dornaika and Davoine [30] use a deterministic registration technique based on Online Appearance Models to estimate head pose. On the other hand, Tian et al. [105] used the silhouette of a head rather than face, which was then converted to grayscale, histogram equalised, and resized to the appropriate resolution. This was

then passed through a Neural Network (NN) with the outputs being one of the following: frontal, near-frontal, side, profile, back or occluded. By using the silhouette the effects of facial expressions were mostly removed, but such representation could be susceptible to ambiguities and could not be used for more fine-grained, or precise, pose definition. Nevertheless the pose information can be extracted from the underlying model - this is especially true for the 3D-based models. Xiao et al. [114] can automatically obtain pose information from the 3D head model which was used to constrain the 2D AAM model. Similarly Xiao et al. [115] developed means of extracting pose information from the cylindrical model.

2.2.2 Expression Categories Extraction

After the extraction of the required features, one may wish to recognise (or classify) facial expression labels based on the given data. Different classifiers can be used for this task. These can be NNs [60, 104, 105, 129, 84, 57], Support Vector Machines (SVMs) [5, 3, 110, 70, 98], BNs [21] and their dynamic counterparts [128, 108, 47, 126], Hidden Markov Models (HMMs) [21, 66, 83, 57] or rule based classifiers [86] and Particle filters [30]. We can form two main categories of approach: static, in which information used for expression analysis is extracted from a single image, and dynamic, in which temporal information is also utilised [106, 85].

Posed versus Spontaneous Expressions

There is a clear physical distinction between posed and spontaneous expressions. Neurological research suggests that they are driven by two different neural pathways in the brain, and that the facial muscles and dynamics are different [85]. For example the spontaneous expressions are driven by different motor pathways, resulting in non-symmetric expressions [106]. Also, spontaneous expressions are context and culture specific, and only small subset of them can be defined across culture or context. Most of the existing work uses data that contains deliberately posed expressions. This data is easier to obtain or generate, but such data rarely exists in a real world scenarios. The use of posed data was partly driven by convenience, but mostly due to the lack of comprehensive and available datasets.

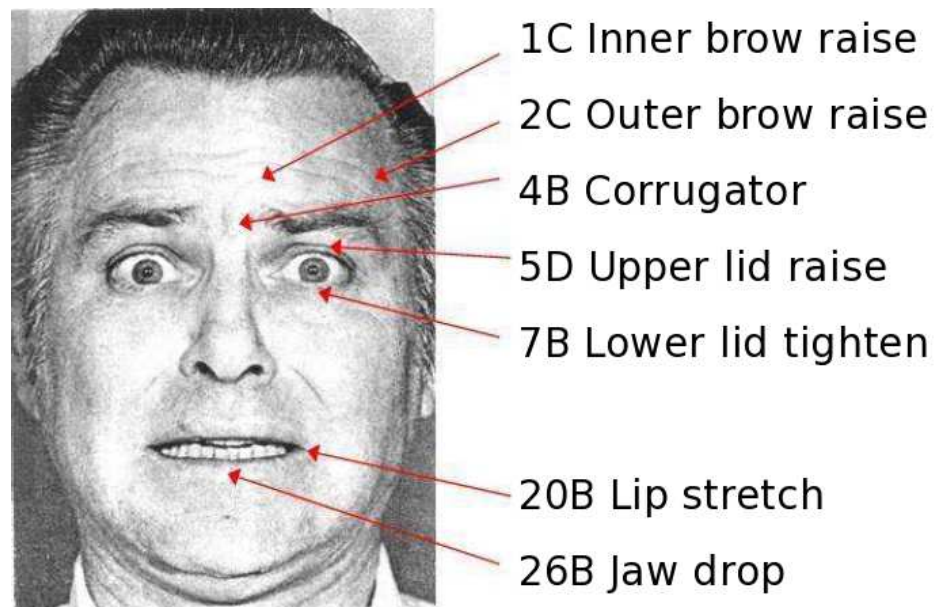


Figure 2.9: The 5-point intensity scale of FACS. From Bartlett et al. [6].

Expression Intensity

We can define expression intensity as a rate of change with respect to some point of reference, usually neutral facial expression. Such a measure is important in the modelling of facial dynamics, and necessary for synthesis and animation of expressions. Pantic and Bartlett [85] point out that there is a coupling between expression decoding accuracy, perceived intensity of the underlying emotional state, and physical intensity of the facial emotion. An expression intensity measure is important for accuracy, and has significant impact on discrimination between posed and spontaneous expressions. As facial expressions rarely convey only one type of emotion, it is also important to be able to detect and extract a combination of displayed emotions and their respective intensities. For basic expressions, such intensities are usually represented as normalised, continuous values within a specified range. The FACS coding system on the other hand uses a 5-point discrete intensity scale (A-F) to describe intensity variation in each of the AUs [106, 85]. Figure 2.9 shows an example of such a scale. Also some of the AU's themselves, combined in groups, can function as a measure of

intensity, e.g. AUs 41-45 can represent intensity change within the eye region [106]. Most of the existing work treats expression classification as a binary state (present or absent), and only a small subset address, or provide the means to define expression intensity as a continuous or multi-level discrete measure [66, 58, 67, 128, 100, 68, 118, 65].

Static Approaches

Kimura and Yachida [58] proposed a method for recognising three kinds of expressions: happiness, anger and surprise. They used potential net motion flow information generated by comparing the expressionless reference frame with the target expression. Then PCA was applied to the extracted vectors to create what they call an emotion space. Intensity and the type of the expression was defined as the distance and direction of a straight line in first two principal components, approximated by the least squares method. This assumes that the expression subspace is constructed in a star-shaped manner with neutral expression at the centre, which might not be true for a larger and non-standard set of expressions. Huang and Huang [53] used a 2D Gaussian mixture model of PCA-represented Action Parameters (APs) space. The PCA space was constructed using 180 location (90 feature PDM) and 13 mouth shape parameters. Combinations of APs defined by Vanger et al. [111] were then used to classify six prototypic expressions. Due to overlaps of each of the expressions with two or more other expressions, in the PCA space, the highest score of the three correlations determined the final expression achieving average recognition rate of 84%. Cohen et al. [21] investigated the use of Bayesian Network Classifiers, focusing on underlying distribution assumptions and design of the network with respect to feature dependency. The underlying feature representation consisted of Motion Units (MUs). In their first experiment they used a Naive-Bayes classifier (Figure 2.10 (a)) and compared the use of the Gaussian and Cauchy distributions. Their findings suggest that the latter performs better, however the independence assumption imposed by the model might not hold true due to correlation between various facial motions during the expressions. The second experiment used a Tree-Augmented-Naive (TAN) Bayes classifier (Figure 2.10 (b)) to automatically learn the dependencies amongst different features in order to overcome the implied feature independence shortcoming of Naive Bayes. They reported recognition rates of 79.36%, 80.05% and 83.31%

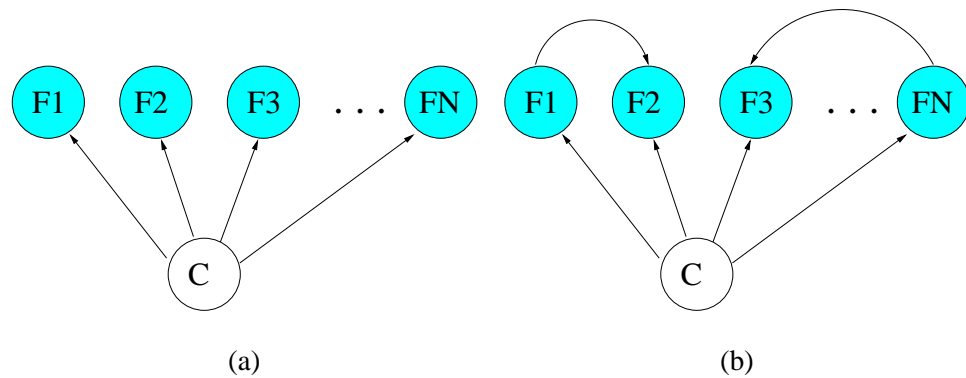


Figure 2.10: Visualisation of Naive Bayes (a) and Tree Augmented Naive Bayes (b) Network of Cohen et al. [21].

for Naive-Bayes Gaussian, Cauchy and TAN classifiers respectively. Neither approach however addresses the issue of modelling expression intensity. Pantic and Rothkrantz [88] employed a rule based forward reasoning classifier to recognise six basic emotions from an AU encoded representation of the expression with 91% success rate. In their following work, Pantic and Rothkrantz [89] extended the previous approach and used two rule-based classifiers to recognise single, or a combination of, 24 AUs in profile view and 22 AUs in frontal view. If both frontal and profile views are available then their system is able to recognise single or combinations of 32 AUs. As the primary focus was on AUs recognition, the logic behind and corresponding set of rules was complex and in the case of [88] was computationally intensive. Littlewort et al. [70] and Bartlett et al. [5] compared the performance of an SVM classifier, AdaBoost and AdaSVM classifiers. A Gabor representation was used, based on patches extracted from the output of the face detector as input to the classifier. For classification a two stage process is performed. Firstly, seven SVM classifiers, one for each emotion class, were used to discriminate each emotion from all others. Then the decision regarding emotion category was made based on the classifier with the maximum margin. Multiple kernels were tested, with Linear and RBF kernels using a unit-width Gaussian performing best. The performance was then compared with an AdaBoost classifier using individual Gabor patches as features. Finally a combination of

AdaBoost selected features were used to provide a reduced representation for the SVM classifier, which showed the best results of all. Expression intensity was calculated by passing a 7-D emotion structure, obtained from the output of the classifier to a CU Animate tool [75], in which each of the expressions was calculated as a weighted combination of morph targets for each emotion. The same approach was adapted to recognise seven upper face AUs by Littlewort et al. [68] and twenty AUs in a non-posed environment by Bartlett et al. [6]. In addition [6] was able to determine AU intensity using the output margin of the SVM classifier. Tian et al. [104] used two (one for the upper and one for the lower face) three-layer NN with one hidden layer to recognise 16 AUs (6 for the upper and 10 for the lower face). According to Tian et al. [106] this is the first system to handle a combination of AUs. Although they have taken the advantage of localised representation, such localisation was performed at the very coarse level. Reilly et al. [94] compared the performance of KPCA with LLE using SVM. They focused only on four lower facial AUs. Rather than trying to estimate intensity they performed classification under varying intensity. The limitation of the above approaches is that they perform recognition without any temporal component of facial expression, which is an important factor in expression recognition [9].

Dynamic Approaches

Eisert and Girod [34] used a 3D model, optical flow and 3D motion equations to produce an FAP description. Similarly Sarris et al. [96] used a 3D model and optical flow to extract appropriate motion parameters using geometric transformations. The extraction process was bootstrapped by previously extracted pose information. In both approaches, there were no attempts to explicitly provide expression labels or intensity, because this information was intrinsically embedded in FAP parameters. Bettinger et al. [12] presented an approach to modelling the dynamics of facial expression. An AAM was used to model the appearance, and image sequences were represented as trajectories in a parameter space. The trajectories were broken into segments, and a variable length Markov Model was used to learn the relationships between those segments. The authors were able to synthesise novel sequences, but did not provide any means for explicit expression labelling or intensity estimation. Otsuka and Ohya [83] used HMM to spot five states corresponding to different contractions of facial

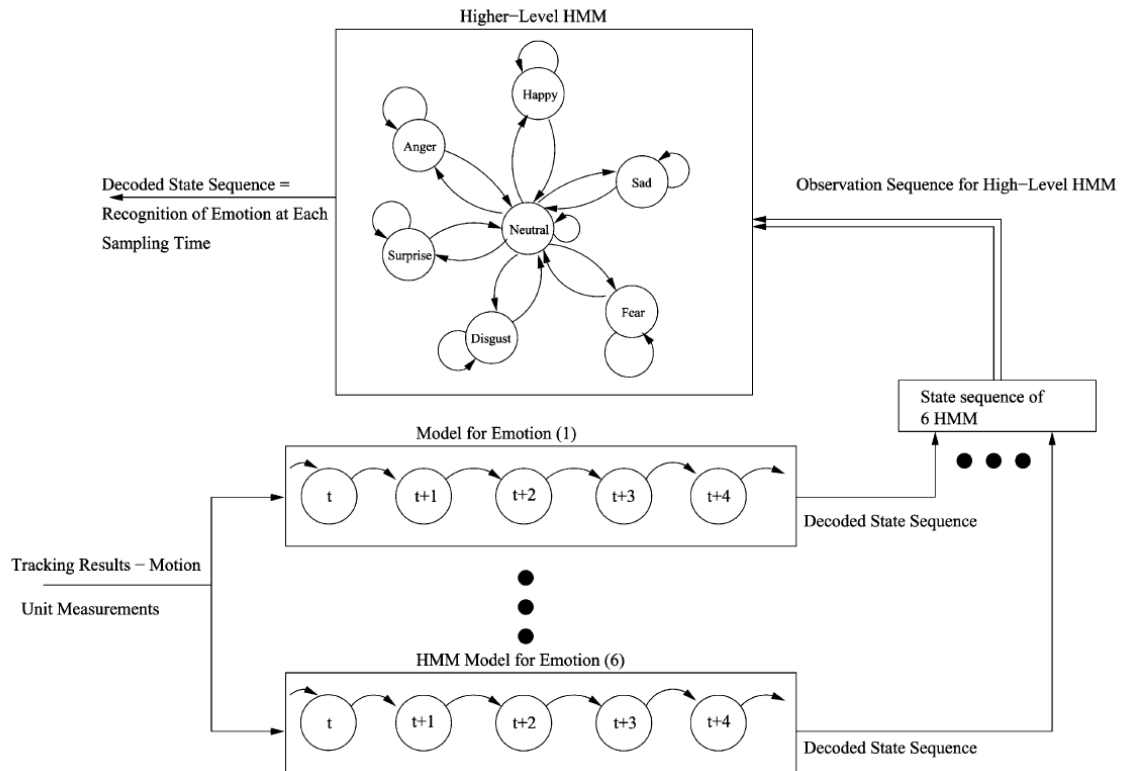


Figure 2.11: Multi level HMM model. From Cohen et al. [21].

muscles: relaxed (neutral), contraction (onset), apex, relaxation (offset). Transition probabilities were conditioned using thresholds to constrain the model and prevent misclassification. No attempts were made to recognise individual, or combination of AUs or intensity associated with them. Lien et al. [67] used a set of HMMs representing individual or combinations of AUs. Once the input expression has been determined the intensity is defined by computing the sum-of-squared differences in PCA space. Cohen et al. [21] used a multi-level HMM for expression classification. The model's higher level section, which defined six basic emotion states such as surprise, disgust, anger, fear, sadness and happiness, was represented as a star shaped model with neutral being in the middle. The lower level consisted of six HMMs, one relating to each of the recognised expressions (Figure 2.11). The lower-level model used MUs as its feature inputs and higher-level model used the

six-state feature vector obtained from the lower-level model. Their approach performed successful segmentation of a video sequence into sections containing different facial expressions. Koelstra and Pantic [61] used GentleBoost for classification and a separate HMM for each of 27 AUs to detect and model their dynamics. For each of the AUs they were able to detect four temporal segments, based on the transitions between neutral, onset, apex and offset, however no fine-grained intensity estimation was conducted. Tian et al. [105] used a three-layer NN with one hidden layer to recognise six prototypic expressions. The network was trained by the standard back-propagation method and used five location features and twelve zone components of mouth shape features as inputs. Dornaika and Davoine [30] used a particle filter to classify six expressions and estimate head pose. Valstar and Pantic [110] proposed a method to recognise 15 AUs individually, or in combination, from frontal-face views. They use SVM classifiers for recognition of an AU and its dynamics (i.e. neutral, onset, apex, offset) based on a set of the best features selected by AdaBoost. Pantic and Patras [86] used a temporal rule-based classifier on profile-view faces to recognise, segment, and model the dynamics of 27 AUs individually, or in combination. Similarly to Valstar and Pantic [110], a particle filter was used to track 15 facial points. Yacoob and Davis [116] used local parametric motion models and a heuristic classifier to discriminate one of six facial expressions. Anderson and McOwan [3] used velocity information obtained from optical flow and an SVM classifier to classify the expressions. Amin et al. [2] employed PCA-based representation of Gabor filters for dimensionality reduction, and fuzzy c-Means clustering to provide best pair-wise match of principal components. Membership of the clusters is mapped to degrees of facial expression intensity. However they only deal with two types of expressions: happiness and surprise, and employ a three-grade intensity scale (less, medium or very). Shan [100] used fuzzy k-Means and a similar three-grade intensity scale, but his approach included all of the six basic expressions. Zhang and Ji [126] used multi-sensory information fusion and Dynamic Bayesian networks to model the temporal behaviour of Action Units and classification of six prototypic expressions. As a follow-up, Zhang and Ji [128] used the same information fusion technique but focused on modelling of temporal changes and intensity variation, and on reduction of inter-personal variations. Lucey et al. [73] used the AAM representation and compared Nearest

Neighbour classifiers based on PCA and LDA subspaces with an SVM classifier in recognition of spontaneous facial expressions using FACS. Their findings suggest that there is no real advantage of NNs over SVM classifiers. The experiments were restricted to one FACS intensity scale (peak). Yang et al. [118] treated intensity estimation as a ranking problem and used RankBoost to model it. The ranking score was used as a direct measure of the intensity, and extended to incorporate expression classification. They also obtained significant performance improvement with RegRankBoost, which introduced L1 norm regularisation into RankBoost. Lee and Xu [65] used the Cascade Neural Network, where structure was learned during a training stage, together with an SVM to estimate facial expression intensity. They experimented with different SVM kernels, and concluded that polynomial kernels performed best. However they only dealt with happy, angry and sad expressions.

2.3 Summary

In this thesis we address the following issues:

1. **Facial feature representation:** Facial feature representation has been widely discussed in the existing literature to date. However the most effective choice and representation of features still remains an open question. In Chapter 3 we explore the concept of *localised representation* for the purpose of modelling the expression space, and present its advantages over the holistic approach. We also introduce a hierarchical model decomposition in order to help produce semantic information using geometric features alone. This produces a compact representation, reduces inter-feature correlations, and to some extent, inter-person variations. We investigate its advantages and disadvantages with respect to appearance and combined representations. Multiple components of such a representation can be utilised for different tasks, such as pose estimation.
2. **Representation of the facial expression subspace:** Correct representation of the facial expression subspace is an important factor for interpretation and modelling. Such a mapping is usual non-trivial and involves projection of higher-dimensional data onto a lower-dimensional

space. In Chapter 3 in order to provide the best representation, we model regions of interest, such as the eyes and mouth, according to their intrinsic functionalities, rather than using a holistic approach that models them according to the basic expression types. For this task we employ a probabilistic framework which is a form of Hierarchical Latent Variable Model that is applied to each of the regions. We show that such a model, employing geometric features alone, removes some of non-linearities from the underlying structure, and is able to model variations using a small set of training samples that generalises well to larger, sample sets.

3. **Pose estimation:** Pose information is crucial for realistic animation. Currently there are few works that provide a unified framework that is capable of modelling the expressions and pose together. Also such pose information can be fed back into the framework to improve expression classification, usually when large pose variations are present. In Chapter 4 we introduce a pose model trained on a sparse set of training samples, which can provide continuous pose estimation and minimise effects of facial expressions.
4. **Semantics formulation:** Being able to provide an appropriate semantic description is necessary to successfully mimic the expression exhibited by a synthetic counterpart. In Chapter 5 we investigate the concept of facial region necessity with respect to expression definition. We modify the commonly used standard set of prototypic expressions to achieve best visual impact, for example we represent the happy expression as both smile and grin states. We present the use of the rule-based classifiers and compare it with BN and holistic approaches in a context of sufficiency for the task at hand without the use of temporal information. Such an approach also provides facial expression intensity estimation. Finally we produce a parametric description of expressions that includes intensity and pose information, that can be applied to a 3D avatar in a morph-target based fashion.

Chapter 3

Hierarchical Feature Representation

Being able to represent facial features in an effective and robust manner is a crucial step for modelling and understanding facial behaviour. One might say it forms a basis for it, as it influences all the steps that will follow.

In this chapter we consider a facial feature representation based on an Active Appearance Model (AAM), but more precisely focus on a hierarchical decomposition of its shape component that is also referred to as the Point Distribution Model (PDM). Although shape models cannot encode information such as skin changes they provide significantly lower feature and hence model dimensionality when compared to appearance models, or combined shape and appearance models. Also removal of the appearance component provides invariance to illumination changes and to some extent to intra-person variations. Hierarchical decomposition further reduces the resulting model dimensionality, aids in the removal of non-linearities caused by large variations, and decreases the number of possible combinations that must be accounted for in order to model the set of required expressions.

We begin by describing how shape and texture are used by the AAM representation. We then follow this by expression manifold representation, where we introduce our hierarchical feature and demonstrate its advantages. Finally in the experiment section we demonstrate how such hierarchical

representation can further reduce overall model dimensionality and analyse its performance.

3.1 Active Appearance Models

Active Appearance Models (AAMs) were introduced by Edwards et al. [33], and have been widely established as a technique for face detection and tracking. They have also appeared in modelling and analysis of facial expressions and their underlying manifolds. AAMs are parametric models which consist of shape and appearance components modelled using Principal Component Analysis (PCA). In addition, the resulting shape and texture models can be combined and modelled under the assumption that both the shape and appearance are linearly related. There is also the alternative approach that treats shape and appearance independently of each other and is referred to as Independent Appearance Models (IAMs). Such decoupling opens up interesting possibilities, by which each of these components can vary independently of the others, but this increases the overall dimensionality of the model. Matthews and Baker [77] improved fitting performance of the AAM by employing an Inverse-Compositional Algorithm in such a decoupled representation.

Gross et al. [46] compared two categories of AAMs: generic and person-specific, where the first tries to model variation of many individuals, and the second focuses on the variation of a single individual. Their empirical evaluation showed that the overall performance of person-specific AAMs in terms of modelling and fitting is better than for generic models. In this chapter, we follow the same line of thought, as we consider that person-specific models are better able to capture the intricate dynamics of facial expressions, provide a better tracking basis, and are more robust under sparse training sample sets.

3.1.1 Shape Component

The shape of any object can be defined by a set of J D -dimensional points. These points may represent boundaries of the given object, or key feature points. The points can take an arbitrary dimension, that is $D \in \mathcal{I}^+$, where \mathcal{I}^+ is positive integer space, although we will focus on 2D spatial coordinates representing positions of the selected landmarks in the image plane. Usually, these coordinates are



Figure 3.1: Selected training samples with overlaid PDM mask.

obtained by manually labelling the training set, but they can also be extracted by using automated labelling procedures, such as the one described by [24]. Figure 3.1 shows some of the training examples with an overlaid shape mask. Let us define vector $\mathbf{s} = \{x_1, y_1, x_2, y_2, \dots, x_J, y_J\}^T$ to represent our shape. Given N training examples, the data set is given by $\mathbf{D} = \{s_1, s_2, \dots, s_N\}$. These vectors form a distribution in a $2J$ dimensional space. To minimise the effects of global transformations, such as rotation or scale, these vectors are aligned in a common co-ordinate frame. This is achieved by Generalised Procrustes Analysis (GPA) [44], and the procedure is as follows:

1. Translate each of the shapes from the training set so that its centre of gravity is at the origin.
2. Choose and scale one of the samples \mathbf{t}_0 as an initial estimate of the mean, such that $\bar{\mathbf{x}}_s = \mathbf{t}_0$ and $|\bar{\mathbf{x}}_s| = 1$.
3. Align all the shapes with the current estimate of the mean using GPA.
4. Re-estimate the mean shape from the aligned shapes.
5. Constrain the current estimate of the mean by aligning it with \mathbf{t}_0 and scaling such that $|\bar{\mathbf{x}}_s| = 1$.

6. Return to step 3 if not converged (convergence is determined to be when there is no further significant change of the mean estimate between iterations).

Once the shapes have been aligned in a common co-ordinate frame, we calculate the mean

$$\bar{\mathbf{x}}_s = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \quad (3.1)$$

and covariance matrix

$$\mathbf{\Sigma}_s = \frac{1}{N} \sum_{n=1}^N [\mathbf{x}_n - \bar{\mathbf{x}}_s][\mathbf{x}_n - \bar{\mathbf{x}}_s]^T \quad (3.2)$$

of the training set. After applying PCA, we obtain eigenvectors \mathbf{u}_i and corresponding eigenvalues λ_i of $\mathbf{\Sigma}_s$. Next the eigenvalues are sorted into decreasing order (i.e., $\lambda_i \geq \lambda_{i+1}$). If \mathbf{U} contains the eigenvectors corresponding to the t largest eigenvalues then any example can be approximated by the following:

$$\mathbf{x}_s \approx \bar{\mathbf{x}}_s + \mathbf{U}\mathbf{b}_s \quad (3.3)$$

where \mathbf{b} is a t dimensional vector such that

$$\mathbf{b}_s = \mathbf{U}^T (\mathbf{x}_s - \bar{\mathbf{x}}_s) \quad (3.4)$$

which defines a set of parameters controlling the model. In our experiments the dimensionality t of the model is determined by retaining 98% of the variation present in the training set. Alternatively, the choice of the number of model parameters can be selected on the model's ability to approximate any of the training samples with pre-defined accuracy. This involves building multiple models, and choosing the one that best satisfies the criteria [24].

By varying the elements of \mathbf{b}_s one can vary the shape \mathbf{x}_s using Equation (3.3). The variance of the i^{th} parameter b_i across the training set is given by λ_i . By constraining b_i , such that $b_i \in [-2.5\sqrt{\lambda_i}, +2.5\sqrt{\lambda_i}]$, we ensure that newly generated shapes are similar to those of our training set. Under the linearity assumption of the shape space, this constraint also defines a Valid Shape Region (VSR). Figure 3.2 shows the effect of varying each of the shape parameters in turn between ± 2.5 standard deviations.

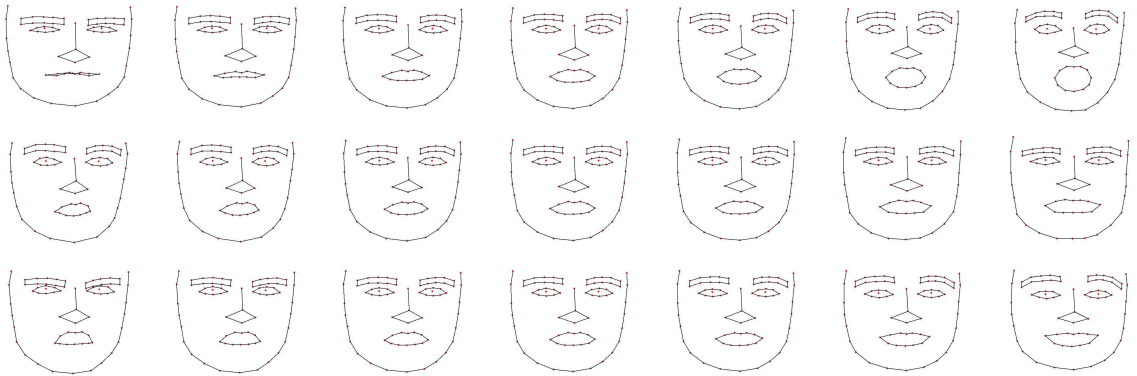


Figure 3.2: Effects of varying the first three (top to bottom) shape parameters in turn between ± 2.5 standard deviation.

3.1.2 Appearance Component

The appearance is defined as a pattern of pixels across an image patch. The patch in the training set is described by the set of points belonging to the corresponding PDM. However the number of pixels in each patch may be different due to external factors, such as scale or head orientation. In order to perform successful analysis of the texture, a coordinate frame has to be found in order to establish correspondence between pixels in the training set.

Firstly, a reference shape is defined, either by selecting a mean shape from the corresponding PDM, or by choosing a sample from the shape training set. Next all the training examples can be morphed to the reference shape to obtain a shape free patch, either by using piece-wise affine morphing [24], or with Thin Plate Splines [14]. Once our shape free patches have been extracted the pixel information is sampled and stored in the vector \mathbf{g}_{im} . Each patch now contains very little texture variation caused by exaggerated expressions and differences in shape. Figure 3.3 shows part of the training image with a shape overlaid on top of it (left image) and the corresponding shape free texture patch (right image). Next, normalisation is performed to minimise the effects of global lighting variation by scaling α and offset β such that

$$\mathbf{g}_t = (\mathbf{g}_{im} - \beta \mathbf{1}) / \alpha \quad (3.5)$$

where $\alpha = \mathbf{g}_{im} \cdot \bar{\mathbf{g}}_t$ and $\beta = (\mathbf{g}_{im} \cdot \mathbf{1}) / n$. This is an iterative process, similar to shape alignment, i.e.

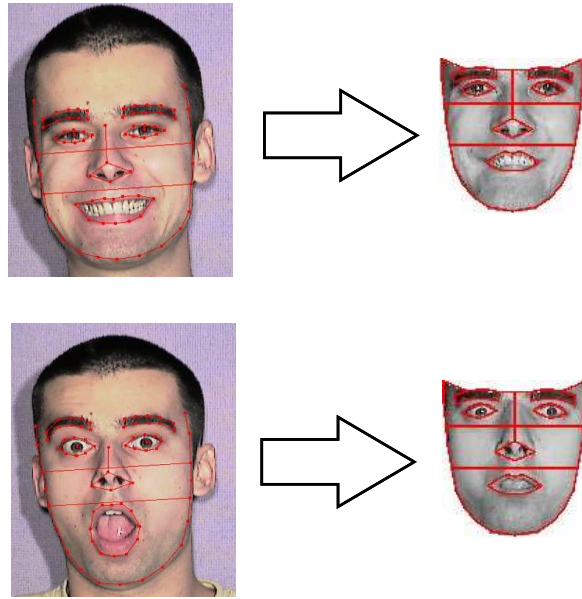


Figure 3.3: Two selected training images (top and bottom) with corresponding shape mask (left) and shape free texture patch (right)

1. Choose one of the samples \mathbf{g}_0 as an initial estimate of the mean, such that $\bar{\mathbf{g}}_t = \mathbf{g}_0$.
2. Align all of the samples with the current estimate of the mean using Equation (3.5).
3. Re-estimate the mean from the aligned samples.
4. Return to step 2 if not converged (convergence is determined to be when there is no significant change of the mean estimate between iterations).

Next the mean

$$\bar{\mathbf{g}}_t = \frac{1}{N} \sum_{n=1}^N \mathbf{g}_n \quad (3.6)$$

and covariance matrix

$$\Sigma_t = \frac{1}{N} \sum_{n=1}^N [\mathbf{g}_n - \bar{\mathbf{g}}_t][\mathbf{g}_n - \bar{\mathbf{g}}_t]^T \quad (3.7)$$

of the training set are calculated. By applying PCA a linear model is given by:

$$\mathbf{g}_t \approx \bar{\mathbf{g}}_t + \mathbf{P}_t \mathbf{b}_t \quad (3.8)$$



Figure 3.4: Effects of varying the first three texture parameters in turn between ± 2.5 standard deviations.

where $\bar{\mathbf{g}}_t$ represents mean of the normalised data, \mathbf{P}_t is the set of eigenvectors corresponding to the largest k eigenvalues which define a set of orthogonal modes of variation, and \mathbf{b}_t is the set of parameters controlling the model. The actual image texture \mathbf{g}_{im} can be generated using previously calculated normalisation parameters using:

$$\mathbf{g}_{im} \approx \alpha(\bar{\mathbf{g}}_t + \mathbf{P}_t \mathbf{b}_t) + \beta \mathbf{1} \quad (3.9)$$

Figure 3.4 shows different texture examples obtained by varying the first three parameters of \mathbf{b}_t in turn by ± 2.5 standard deviations.

3.1.3 Shape and Appearance Combined

The shape and texture can be represented by the parameter vectors \mathbf{b}_s and \mathbf{b}_t respectively. Assuming that there is correlation between the two, the combined sample vector is defined as follows:

$$\mathbf{b}_a = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_t \end{pmatrix} = \begin{pmatrix} \mathbf{W}_s \mathbf{P}_s^T (\mathbf{x}_s - \bar{\mathbf{x}}_s) \\ \mathbf{P}_t^T (\mathbf{g}_t - \bar{\mathbf{g}}_t) \end{pmatrix} \quad (3.10)$$

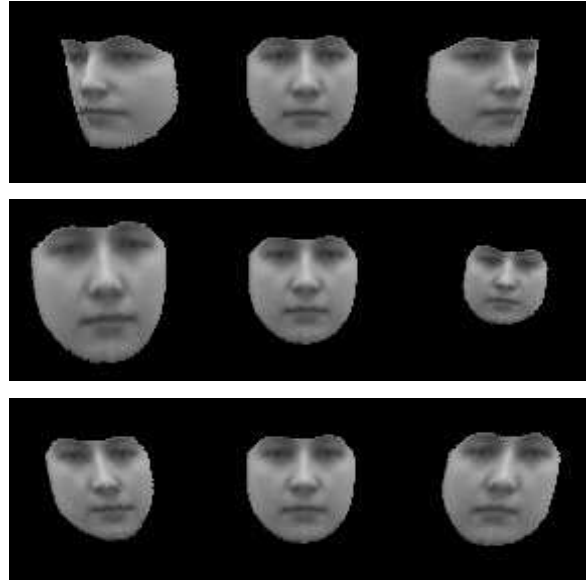


Figure 3.5: Shape variation obtained by varying the first three appearance parameters in turn between ± 2.5 standard deviations.

were \mathbf{W}_s is a weight matrix with diagonal elements to compensate for difference in units between shape and appearance. Such a matrix can be defined as

$$\mathbf{W}_s = r\mathbf{I} \quad (3.11)$$

where r^2 is the ratio of total intensity variation to total shape variation in the normalised frames. Next PCA is applied to the combined sample vectors, yielding the following model

$$\mathbf{b}_a = \mathbf{P}_a \mathbf{a} \quad (3.12)$$

where \mathbf{P}_a are the eigenvectors and \mathbf{a} is a vector controlling the appearance parameters. Figure 3.5 shows different examples obtained by varying the first three parameters of \mathbf{a} in turn by ± 2.5 standard deviations.

3.2 Expression Manifold Representation

Given a combined shape and appearance model, the main challenge we investigate here is how to provide a meaningful representation of the expressions and an accurate estimate of this representations underlying manifold. This expression manifold will help with the modelling of each expression's dynamics and in turn aid resulting parameterisation. To demonstrate the range of expressions we want to model we have selected those that provide the most visual impact. Although based on the set of universal expressions (happiness, sadness, surprise, fear, anger and disgust) defined by Ekman [35], it differs in two following ways: firstly, we divide happiness into two separate smile and grin expressions, and secondly we do not distinguish between surprise and fear and treat them both as a single, or joint, label. The former choice is influenced by similar expression categorisations that have been used in the gaming and on-line communities in the form of emoticons, where smile and grin represent two distinct states. A similar distinction has been used in facial animation to emphasise different levels of happiness [42]. The latter choice was influenced by the analysis of our training set, and observation of a close similarity between fear and surprise expressions: they only differed in the underlying intensity.

We have found that because AAMs, and their respective shape and appearance components are modelled using PCA, which is based on linear statistics, each of the modes can only vary along a straight line, and non-linear modelling is achieved by a combination of two or more modes [49]. Although Hong et al. [51] and Cho et al. [18] have reported success using PCA to represent the expression manifold, they only considered three expressions: disgust, happiness and surprise. This is understandable, as the number of variations present in their dataset was relatively small. When one considers more varied expression changes, PCA is not very well suited to manifold representation as it cannot reliably capture, or separate, subtle intra-feature variations and determine the true degrees of freedom, when large variations, due to expression or pose changes, occur [17].

Even in datasets where variations due to inter-person differences are not present, such a representation still does not yield a compact and clearly defined grouping with respect to the predefined labels in the resulting subspace. Figure 3.6 shows the first two (left column) and second two (right

column) principal components for holistic shape, appearance and combined shape and appearance (top to bottom) respectively, of the AAM trained using 886 labelled samples from a person-specific dataset. To overcome the limitation of PCA based techniques, various approaches have been proposed. Chang et al. [17] applied a modified Lipschitz embedding to the training set, consisting of six reference sets for each of the modelled expressions. Shan et al. [97] introduced Locality Preserving Projections to model the manifold of expression space. Cootes et al. [23] used separate AAMs for profile, half-profile and frontal views, although their primary motivation was to account for viewpoint variation. A similar approach was undertaken by Moghaddam and Pentland [79].

We propose a different solution. Rather than treating facial expression as a holistic entity, and modelling it as such, we define it as the combination of the most salient and intuitive facial regions. This is because our experiments [121] suggest that for particular facial expressions, only certain regions convey the most relevant information, and the contribution of others is marginal. For example when we grin, mainly the lower part of the face; the mouth shape together with possible skin creases around the nose area; contains the relevant information, and when we are surprised, relevant information is mostly conveyed through mouth shape and widening of the eyes. Figure 3.7 shows the different activation areas for two different types of expressions, obtained by motion differencing of selected frames with an initial (neutral expression) frame. We can see that for the grin expression (a) the motion is mostly concentrated around the mouth and nose areas, and for the surprised expression (b) concentration falls in the mouth and eye regions. Such a region driven representation has been utilised in the context of computer generated facial animation [42, 63] and also has been investigated in psychophysical experiments and found to be the most descriptive and sufficient in the context of facial expression recognition [27].

3.2.1 Hierarchical Representation

In our new approach we define a hierarchical representation of features in terms of subcomponents based on selected regions of the face. The subcomponents are defined as follows:

- The jaw outline, nose, centres of the eyes and mouth form the root of our hierarchy.

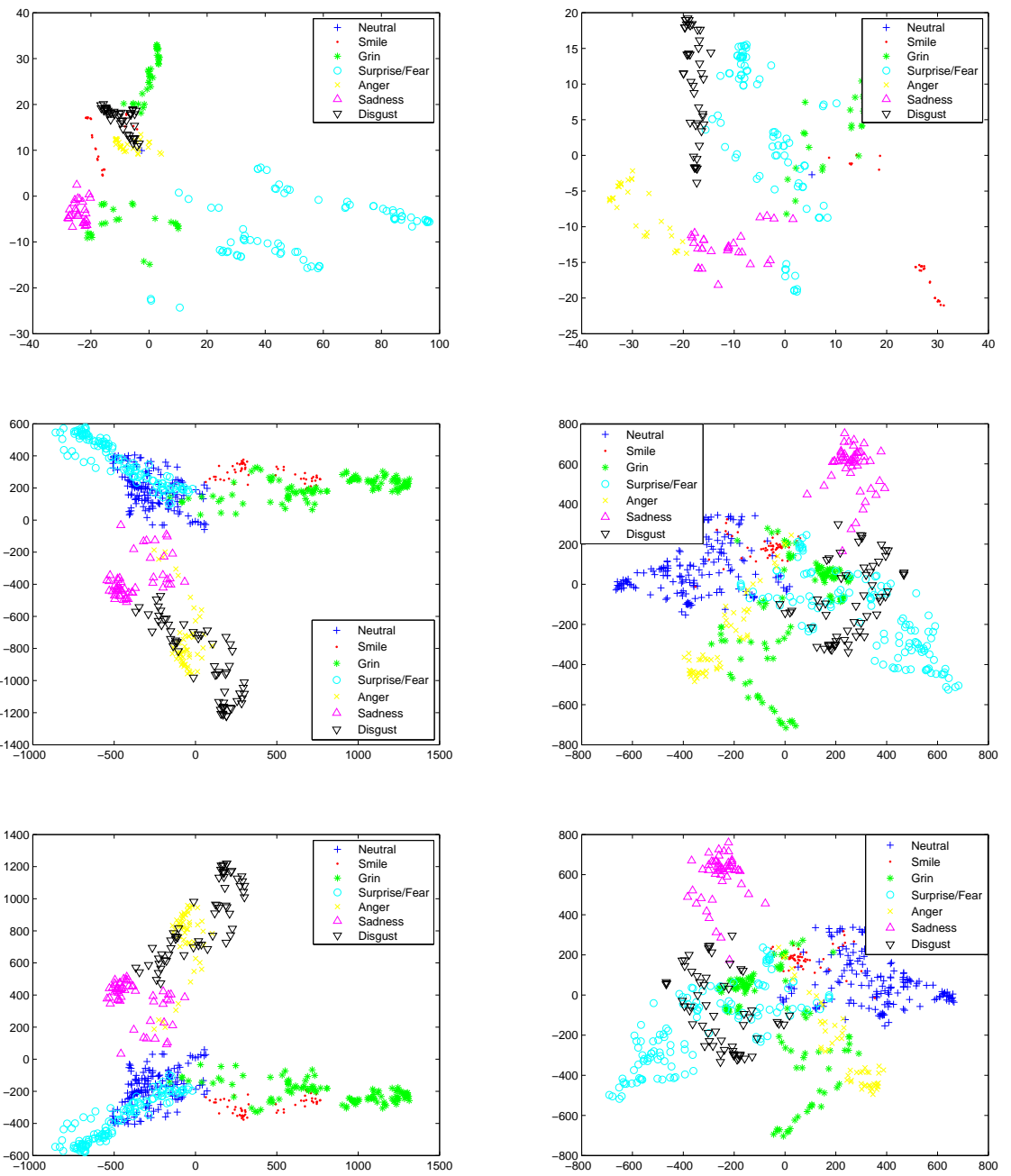


Figure 3.6: First (left column) and second (right column) are the two principal components of the training set that have been projected onto shape (top row), texture (middle) and combined shape and texture (bottom) in the expression space.

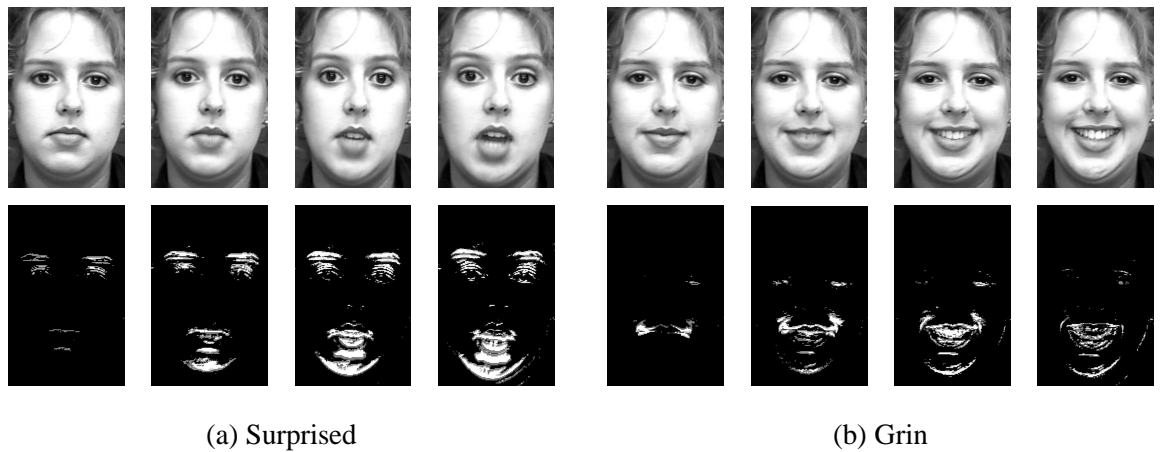


Figure 3.7: Original image (top rows in (a) and (b)), and the regions of the image exhibiting motion during expressions (bottom rows) such as surprise (a) and grin (b).

- With the leaves, or children, being used to model the eye eyebrow pairs and the mouth.

Figure 3.8 shows an example of such a decomposition. The leaves are used for expression modelling,

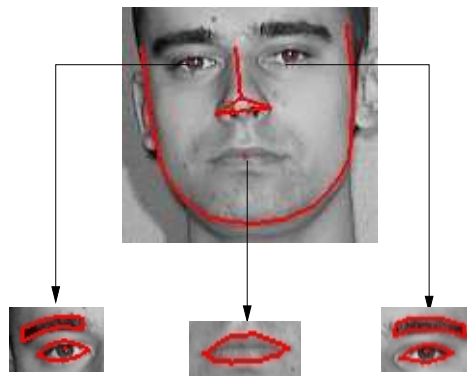


Figure 3.8: Structure of our hierarchy. The top row corresponds to the highest point in the hierarchy (root), the middle row corresponds to the leaves.

and the root component is utilised for estimating pose (which we will investigate in Chapter 4). Localised representation, based on similar face regions have been used by Padgett and Cottrell [84], however their motivation was different and focused entirely on face detection. The advantages of

localised compared to holistic representation are:

- Firstly of all, each of the expressions is defined in a more intuitive and quantitative way, for example, as a combination of intrinsic functionalities of each of the subcomponents - this form of expression implied facial feature independence has been exploited by [28].
- Secondly, such a representation allows us to account for similar expressions (smile with eyes open, or smile with eyes closed) with a much smaller training set. Figure 3.9 demonstrates four expressions, which for holistic representation would all have to be included in the training set in order to be successfully modelled. For hierarchical representation inclusion only of the first two is necessary: the others will be automatically accounted for.



Figure 3.9: Example expressions that would have to be included in the training set if holistic model was used. In case of hierarchical model only two are necessary.

- Thirdly, we remove the explicit need to model, or account for, correlation between those facial parts at this stage. Also non-linearities resulting from variations caused by large expression changes are minimised. We can also discard some information that might be either unnecessary, or in the worst case actually diluting the true nature of the data.
- Finally, each of the hierarchical components can be modelled according to its intrinsic functionality, for example, eye components could represent wide open, neutral or squint. This allows the capture of more localised variations, and more accurate dynamics. Also this provides

the ability to re-formulate, or learn, the final expression states as a combination of component states, or even represent states of the subcomponents in isolation.

We choose only the shape component for our hierarchical representation. The shape is individually independent given appropriate normalisation, hence can be efficiently utilised to capture manifolds of the facial expressions. Texture information was not selected because it is susceptible to external factors such as illumination or identity changes. Although shape alone is hardly sufficient to represent subtle skin variations such as wrinkles or furrows, it has been shown to perform as well as texture in some circumstances [85]. The use of the shape component on its own has also been adopted by Huang and Huang [53] and Chang et al. [16, 17] to represent and model facial expression subspaces.

A hierarchical approach can also be extended to the underlying data representation layer. This is especially useful in cases where data contains multiple classes, whose variation we wish to capture and represent. Instead of using PCA we adopt Hierarchical Latent Variable Model (HLVM) of [13]. This approach provides several advantages in terms of clustering, density modelling, and dimensionality reduction. The main shortcoming of PCA is the lack of a probability density framework [107]. By incorporating this into Bayesian frameworks, it would become possible to model class based densities, and to provide interoperability should there be any missing values. Also, most importantly, non-linear variations could be represented by a collection of localised linear models. Finally, all of the model parameters can be determined in a maximum likelihood framework, where the partitioning of the data and the calculation of respective principal components are obtained automatically as the likelihood is maximised.

3.3 Hierarchical Latent Variable Model

Bishop and Tipping [13] introduced the HLVM, which is an extension of the Mixture of Probabilistic Principal Component Analysers [107]. Figure 3.10 demonstrates the concept of such a hierarchy, where each of the levels successively defines a more refined and detailed representation of the data. For a given data set $\{\mathbf{t}_i\}$ where $i \in \{1, \dots, N\}$ and each element of \mathbf{t}_i has dimensionality d , and where we have a single latent variable model defining the root of the hierarchy, the linear mapping onto a q

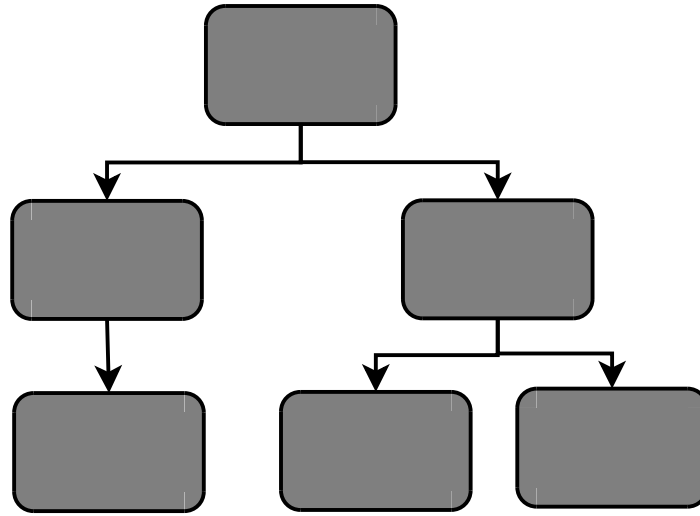


Figure 3.10: Conceptual visualisation of the Hierarchical Latent Variable Model. The top level corresponds to a single latent variable model, and subsequent levels correspond to fine-grained mixtures of them.

dimensional space \mathbf{x} is given as:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \quad (3.13)$$

where \mathbf{W} is a factor loading matrix, $\boldsymbol{\mu}$ is mean and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ is a noise process. For a given \mathbf{x} the probability distribution over \mathbf{t} is:

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \quad (3.14)$$

where $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The probability density function of such a model is defined as:

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (3.15)$$

which is also Gaussian, such that:

$$p(\mathbf{t}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \quad (3.16)$$

where the $d \times d$ model covariance matrix $\mathbf{C} = \sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^T$. The posterior distribution can be defined as:

$$p(\mathbf{x}|\mathbf{t}) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{t} - \boldsymbol{\mu}), \sigma^2 \mathbf{M}^{-1}) \quad (3.17)$$

where the $q \times q$ posterior covariance matrix is $\mathbf{M} = \sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W}$.

Tipping and Bishop [107] showed that the maximum likelihood solution for parameters $\boldsymbol{\mu}$, \mathbf{W} and σ^2 is:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N \mathbf{t}_n \quad (3.18)$$

$$\mathbf{W}_{ML} = \mathbf{U}_q (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I})^{\frac{1}{2}} \mathbf{R} \quad (3.19)$$

$$\sigma_{ML}^2 = \frac{1}{d-q} \sum_{j=q+1}^d \lambda_j \quad (3.20)$$

$$(3.21)$$

where \mathbf{U}_q are the eigenvectors and $\boldsymbol{\Lambda}_q$ are the eigenvalues of the sample covariance matrix \mathbf{S} of the observed values:

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N (\mathbf{t}_i - \boldsymbol{\mu})(\mathbf{t}_i - \boldsymbol{\mu})^T \quad (3.22)$$

A reduced dimensionality transformation of the data point \mathbf{t}_n is given by:

$$\hat{\mathbf{x}}_n = \mathbf{W}_{ML}^T (\mathbf{t}_n - \boldsymbol{\mu}_{ML}) \quad (3.23)$$

and its optimal reconstruction \mathbf{t}_n is given by:

$$\mathbf{t}_n \approx \mathbf{W}_{ML} (\mathbf{W}_{ML}^T \mathbf{W}_{ML})^{-1} \hat{\mathbf{x}}_n + \boldsymbol{\mu}_{ML} \quad (3.24)$$

This can be extended to a mixture of such models, hence defining the second level of the hierarchy.

The mixture density model is then given as follows:

$$p(\mathbf{t}) = \sum_{i=1}^M \pi_i p(\mathbf{t}|i) \quad (3.25)$$

where M defines the number of components in the mixture, and π_i are the mixing coefficients corresponding to the mixture components $p(\mathbf{t}|i)$. Each of the mixture components is a latent variable model, so the model is defined in terms of parameters π_i , $\boldsymbol{\mu}_i$, \mathbf{W}_i and σ_i^2 . To obtain the parameters, Tipping and Bishop [107] show that for a given posterior responsibility of component i , that is generating data point \mathbf{t}_n , we can express:

$$R_{ni} = p(i|\mathbf{t}_n) = \frac{p(\mathbf{t}_n|i)\pi_i}{p(\mathbf{t}_n)} \quad (3.26)$$

and an iterative, two stage EM algorithm produces the following parameter updates:

$$\tilde{\pi}_i = \frac{1}{N} \sum_{n=1}^N R_{ni} \quad (3.27)$$

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{n=1}^N R_{ni} \mathbf{t}_n}{\sum_{n=1}^N R_{ni}} \quad (3.28)$$

$$\tilde{\mathbf{W}}_i = \mathbf{S}_i \mathbf{W}_i (\boldsymbol{\sigma}_i^2 \mathbf{I} + \mathbf{M}_i^{-1} \mathbf{W}_i^T \mathbf{S}_i \mathbf{W}_i)^{-1} \quad (3.29)$$

$$\tilde{\sigma}_i^2 = \frac{1}{d} \text{Tr}(\mathbf{S}_i - \mathbf{S}_i \mathbf{W}_i \mathbf{M}_i^{-1} \tilde{\mathbf{W}}_i^T) \quad (3.30)$$

$$(3.31)$$

where \mathbf{S}_i is defined as the local responsibility weighted covariance matrix:

$$\mathbf{S}_i = \frac{1}{\tilde{\pi}_i N} \sum_{n=1}^N R_{ni} (\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_i)(\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_i)^T \quad (3.32)$$

Given the definition of the second level of the hierarchy, this can be extended further, where each of the M components can be decomposed into O_i subcomponents in the lower levels of the hierarchy.

For each such level, the density function is given by:

$$p(\mathbf{t}) = \sum_{i=0}^M \pi_i \sum_{j \in O_i} \pi_{j|i} p(\mathbf{t}|i, j) \quad (3.33)$$

where $p(\mathbf{t}|i, j)$ defines independent latent variable models, $\pi_{j|i}$ defines mixing coefficients for each i , and $\sum_j \pi_{j|i} = 1$. Each of the given levels corresponds to a generative model, with child levels providing more detailed and refined information. Now, the posterior responsibilities for the component i, j generating a data point \mathbf{t}_n is:

$$R_{ni,j} = R_{ni} R_{ni|j} = \frac{\pi_{j|i} p(\mathbf{t}_n|i, j)}{\sum_{j'} \pi_{j'|i} p(\mathbf{t}_n|i, j')} \quad (3.34)$$

where $\sum_{j \in O_i} R_{ni,j} = R_{ni}$, gives rise to the following update equations:

$$\tilde{\pi}_{j|i} = \frac{\sum_{n=1}^N R_{ni,j}}{\sum_{n=1}^N R_{ni}} \quad (3.35)$$

$$\tilde{\boldsymbol{\mu}}_{i,j} = \frac{\sum_{n=1}^N R_{ni,j} \mathbf{t}_n}{\sum_{n=1}^N R_{ni,j}} \quad (3.36)$$

$$\tilde{\mathbf{W}}_{i,j} = \mathbf{S}_{i,j} \mathbf{W}_{i,j} (\boldsymbol{\sigma}_{i,j}^2 \mathbf{I} + \mathbf{M}_{i,j}^{-1} \mathbf{W}_{i,j}^T \mathbf{S}_{i,j} \mathbf{W}_{i,j})^{-1} \quad (3.37)$$

$$\tilde{\sigma}_{i,j}^2 = \frac{1}{d} \text{Tr}(\mathbf{S}_{i,j} - \mathbf{S}_{i,j} \mathbf{W}_{i,j} \mathbf{M}_{i,j}^{-1} \tilde{\mathbf{W}}_{i,j}^T) \quad (3.38)$$

where again $\mathbf{S}_{i,j}$ is defined in a responsibility weighted covariance matrix:

$$\mathbf{S}_{i,j} = \frac{1}{\sum_{n=1}^N R_{ni,j}} \sum_{n=1}^N R_{ni,j} (\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_{i,j})(\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_{i,j})^T \quad (3.39)$$

3.4 Experiment

Our main training set consists of 1300 640×480 colour images and manually labelled shapes (consisting of 74 landmarks), which include the following six expressions: neutral, smile, grin, sadness, anger, surprise/fear and large variations in pose. Figure 3.1 shows selected samples from the dataset. Although our focus is mainly on person specific expression parameterisation, we also use selected samples from the Cohn-Kanade facial database [20] which contains 486 sequences from 97 subjects. This ensures wider variation of expressions being modelled, accounts for unpredictability in continuously changing facial motion, and provides a better generalisation of the model. For testing of the model we used two sequences **T1** and **T2** consisting of 1536 and 1541 frames respectively, which were tracked using an AAM based tracker and contained a variety of modelled expressions. Those sequences included a substantial amount of noise caused by misalignment failures in the tracking procedure. Although good results were obtained using the training set, or parts of it, we felt that using output from the tracker provided a more realistic test-bed. Figure 3.11 demonstrates some of the correctly (top) and incorrectly (bottom) tracked expressions.

Rather than modelling each of the states as belonging to a set of six classes of prototypic expressions, we investigate each of the hierarchical components with respect to their intrinsic functional labels. For eye components they are neutral, open and squint. For mouth component they are neutral, anger, sad, smile, grin and wide open. In Chapter 5 we present a method which based on these intrinsic functionalities will produce the final expression labels. We define three models, one corresponding to each of the eyes and one to the mouth area. Our implementation of HLVM is based on the PhiVis toolbox of [13].

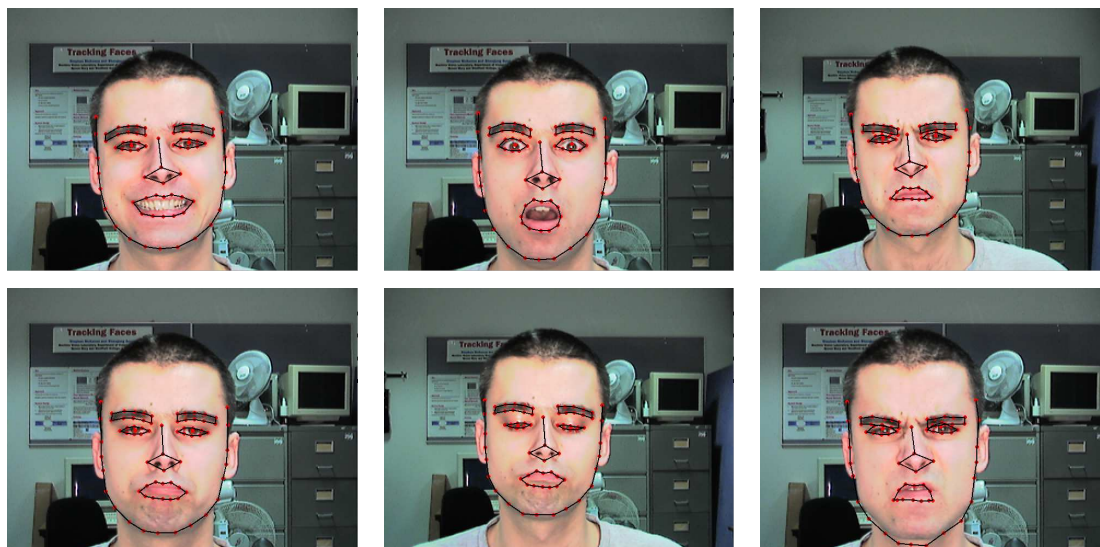


Figure 3.11: Sample of correctly (top) and incorrectly (bottom) tracked expressions.

3.4.1 Eye Models

To build models for both left and right eyes we use a training set of 288 samples from a person specific training dataset, and 114 samples from the Cohn-Kanade dataset, where each of the samples consisted of 19 facial landmarks. Those landmarks were obtained by sub-sampling the holistic shape representation according to the defined hierarchical decomposition rules in Section 3.2.1. The proportions of intrinsic eye states present in the training set are listed in Table 3.1.

Neutral	Open	Squint
25%	33%	42%

Table 3.1: Proportions of intrinsic eye states present in the training set.

We injected some artificial variation by duplicating the training set and perturbing each of the duplicated samples \mathbf{t}_o by a vector \mathbf{v} whose elements were drawn from uniform distribution and scaled such that $\mathbf{v} \in [-1, 1]$. The displaced sample \mathbf{t}_d was given by $\mathbf{t}_d = \mathbf{t}_o + \mathbf{v} * const$ where $const$ determined

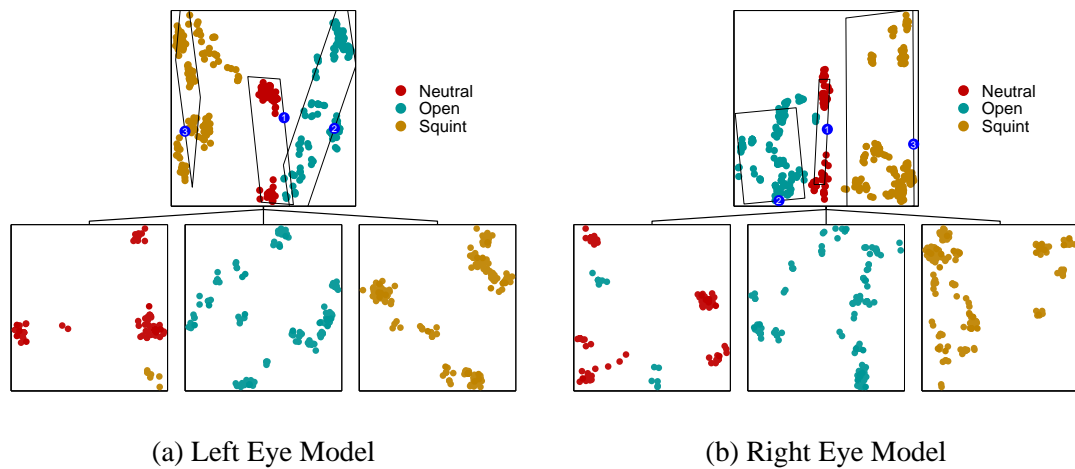


Figure 3.12: Visualisation of the hierarchical clustering in the eye space for (a) left and (b) right eye models. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy. The numbered clusters in the top level correspond to the order of the cluster frames in the lower levels.

the severity of the displacement. This is analogous to the small shifts of landmarks introduced during the labelling process or by tracking errors. A similar concept was used by Cootes and Taylor [24], but their primary motivation was to add extra unseen variation in order to enhance the flexibility of the model, rather than to reinforce the existing variations. The artificially perturbed samples were appended to the original training set. We chose a supervised approach to building the model, where the initial cluster centre positions used to initialise the EM algorithm were selected interactively for each of the levels of the hierarchy of the model. Although this does not yield an automated process, it allows discovery of the internal data structure and creation of the model based on it. Alternative methods such as K-Means, or taking an average value of the data according to the assigned labels, could also be used as a starting points for the EM algorithm. At each level of the hierarchy convergence was assumed after a few iterations. Figure 3.12 shows a visualisation of the resulting models for the left eye (a) and right eye (b). Because the number of modelled classes was small, a two level hierarchy was sufficient for representing the model.

As both eye shapes are nearly symmetric, we can use a single, unified model. This would result

in a smaller number of models being needed to represent the expressions, and reduce the number of overall parameters. By taking one of the data sets belonging to a particular eye, we can mirror the samples along the vertical axis, and after applying the appropriate normalisations, both datasets can be treated as a unified single eye data set. We have to acknowledge that there will always be small discrepancies between the two shapes, mainly due to the fact that the face is not exactly symmetrical and the range of motions exhibited is not the same either. However such an approximation is desirable and can be used to create a model that will capture the required variations in a unified manner.

To that end, we have taken samples from the training set for the right eye, mirrored them along the y-axis and aligned them. Then those samples were used to validate the left eye model. A similar procedure was repeated whereby samples from the left eye were taken, mirrored along the y-axis, aligned and used to validate the right eye model. Table 3.2 shows the resulting 3-class confusion matrices for the left eye model (a) and right eye model (b).

	Neutral	Open	Squint		Neutral	Open	Squint
Neutral	96.64%	3.36	0	Neutral	87.1%	0	12.9%
Open	0	94%	6%	Open	23.3%	71.9%	4.8%
Squint	20.4%	0	79.6%	Squint	0	0	100%

(a) Left Eye Model

(b) Right Eye Model

Table 3.2: Confusion matrices of the left eye model (a) and the right eye model (b).

Given the promising results of models built using data from a single eye, which were cross-validated with the data from the other eye, we used samples from both eyes to train the combined eye model. Once mirrored, aligned and combined, the training set contained 804 samples. Given the definition of a hierarchical latent variable model in the previous section, a visualisation of the combined eye model is shown in Figure 3.13. Compared to the single eye models, this model now contains clusters corresponding to each of the intrinsic labels for each of the eyes. In the case of

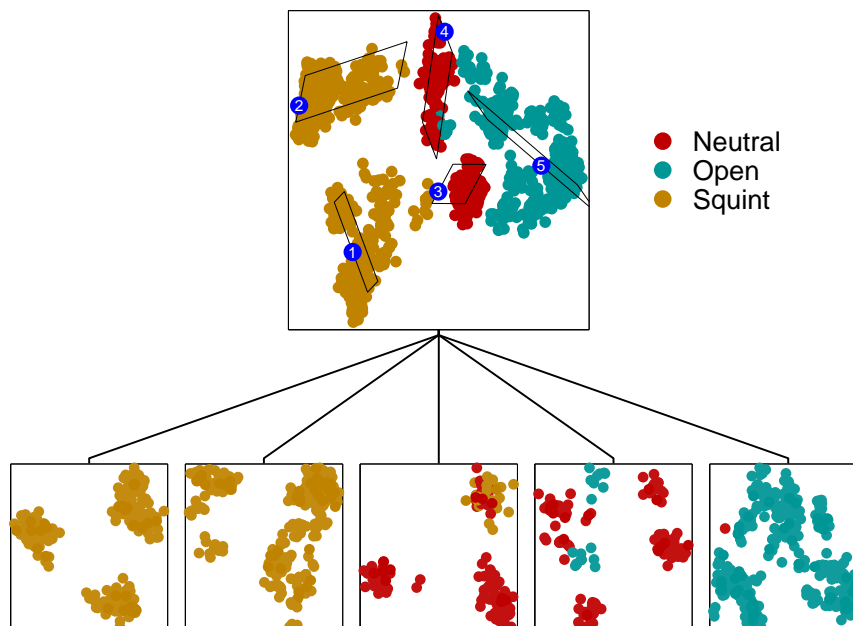


Figure 3.13: Hierarchical clustering in the eye space. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy.

both single and combined eye models, the dimensionality of the latent space was set to $q = 3$. This number was selected experimentally, and chosen to maximise the classification accuracy.

To evaluate the performance of the combined eye model and find out how well it generalises to larger datasets containing unseen samples, we tested it using our two test sequences: **T1** and **T2**. The selected proportions of each of the intrinsic eye states are listed in Table 3.3.

	Neutral	Open	Squint
T1	35%	30%	35%
T2	37%	35%	28%

Table 3.3: Proportions of intrinsic eye states for **T1** and **T2** test sequences.

The classification of intrinsic functionalities was performed by evaluating class conditional probabilities and choosing the class, or label with the highest value. For a given eye shape \mathbf{t}_{eye} this is given

by:

$$j = \arg \max p(j|\mathbf{t}_{eye}) \quad (3.40)$$

where $p(j|\mathbf{t}_{eye}) = \frac{p(\mathbf{t}_{eye}|j)p(j)}{p(\mathbf{t}_{eye})}$. The results were compared with manually assigned ground truth labels.

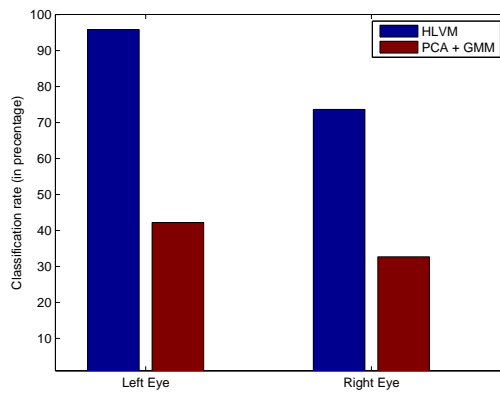
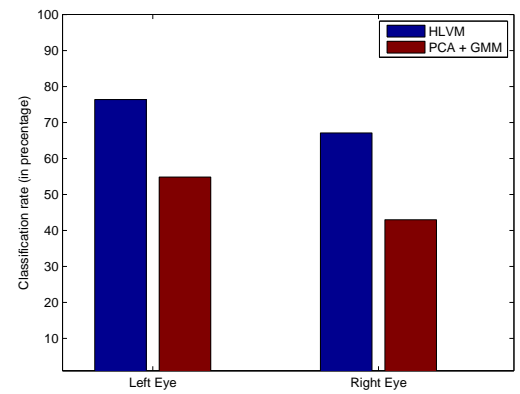
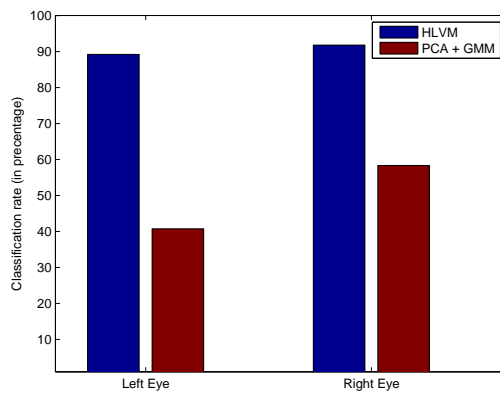
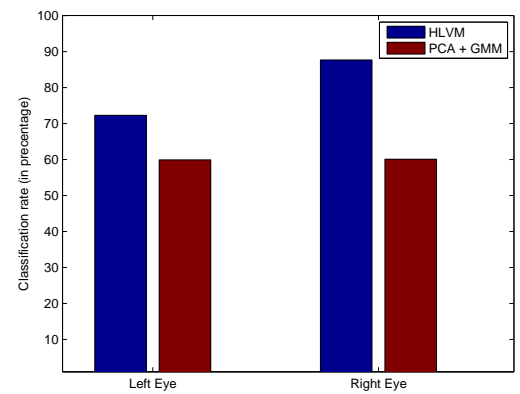
Table 3.4 shows the 3-class confusion matrix of the eye model for test sequence **T1** (a) and test sequence **T2** (b).

	Neutral	Open	Squint		Neutral	Open	Squint
Neutral	89.40%	2.87%	7.73%	Neutral	81.32%	5.34%	13.34%
Open	11.69%	88.31%	0%	Open	32.03%	57.83%	10.14%
Squint	0.28%	0%	99.72%	Squint	1.16%	0%	98.84%

(a)
(b)

Table 3.4: Confusion matrix of the overall eye state classification for test sequence **T1** (a) and **T2** (b) for the combined eye model.

Next we compared our method with that used by Huang and Huang [53] and Liu et al. [71] in which a Gaussian Mixture Model (GMM) was fitted over the existing PCA subspace. We used exactly the same training and validation datasets to train and validate the relevant models. First we compared the performance of our method with the PCA+GMM approach for separate models built for each of the eyes. Figure 3.14 shows the overall results of the experiment. This confirms the advantage of the HLVM approach over PCA+GMM one. We can also see that the right eye model does not perform as well as the left eye model. Next we repeated the same procedure but used the combined eye model. Figure 3.15 shows the overall results of the test. Interestingly for **T1**, the HLVM was able to compensate for poorer performance in the right eye, producing very similar scores for both eyes. This is due to the HLVM ability to capture and represent the local variations more accurately.

(a) Test sequence **T1**(b) Test sequence **T2**Figure 3.14: Classification results for test sequence **T1** (a) and test sequence **T2** (b) using separate eye models for each of the eyes.(a) Test sequence **T1**(b) Test Sequence **T2**Figure 3.15: Classification results for test sequence **T1** (a) and test sequence **T2** (b) using the single eye model.

3.4.2 Mouth Model

For the mouth model we followed the same procedure as with the eye model. We have used 527 training samples from a person specific training dataset and 114 samples from the Cohn-Kanade dataset. Each of the samples consisted of 26 facial landmarks, and those landmarks were obtained by sub-sampling the holistic shape representation according to the defined hierarchical decomposition rules in Section 3.2.1. The proportions of intrinsic mouth states present in the training set are listed in Table 3.5.

Neutral	Smile	Grin	Anger	Disgust	Sad	Open
13%	19%	20%	14%	10%	9%	15%

Table 3.5: Proportions of intrinsic mouth states present in the training set.

Artificial variation was injected in the same way as for the eye models. Similarly, an interactive model creation process was adopted, where at each level of the hierarchy the convergence was assumed after few iterations. The dimensionality of the latent space was set to $q = 3$. Again, this number was selected experimentally in order to maximise the classification accuracy. Figure 3.16 shows a visualisation of the resulting mouth model. To evaluate the performance of the mouth model and to see how well it generalises to larger datasets containing unseen samples, we performed similar test to the eye model, again using our two test sequences **T1** and **T2**. The selected proportions of each of the intrinsic mouth states are listed in Table 3.6.

	Neutral	Smile	Grin	Anger	Open	Sad
T1	14%	16%	18%	21%	12%	19%
T2	16%	22%	13%	19%	5%	25%

Table 3.6: Proportions of intrinsic mouth states for **T1** and **T2** test sequences.

The classification of intrinsic functionalities was performed by evaluating class conditional probabilities and choosing the label with the highest value. For a given mouth shape \mathbf{t}_{mouth} this was given

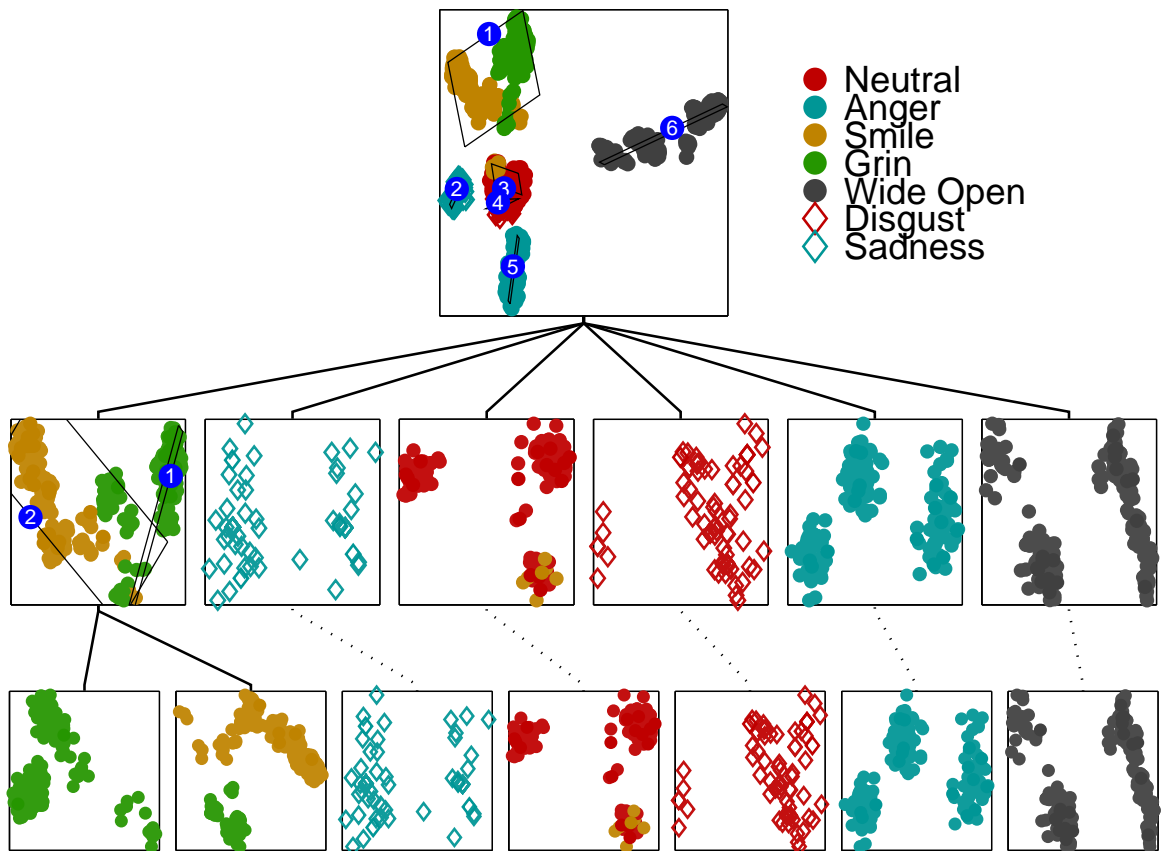


Figure 3.16: Hierarchical clustering in the mouth space. Colours correspond to the intrinsic functionalities of the components. Each of the rows depicts a level in the hierarchy.

by:

$$j = \operatorname{argmax} p(j|\mathbf{t}_{\text{mouth}}) \quad (3.41)$$

where $p(j|\mathbf{t}_{\text{mouth}}) = \frac{p(\mathbf{t}_{\text{mouth}}|j)p(j)}{p(\mathbf{t}_{\text{mouth}})}$. The results were compared with manually assigned ground truth labels. Table 3.7 shows the 6-class confusion matrix of the mouth model for test sequence **T1** (a) and test sequence **T2** (b).

As before, we compared our method with that used by Huang and Huang [53] and Liu et al. [71] in which a Gaussian Mixture Model (GMM) was fitted over the existing PCA subspace. Again, the same training and validation datasets were used to train and validate both models, and Figure 3.17 shows the overall results. The performance of the mouth model is much lower than that of eye

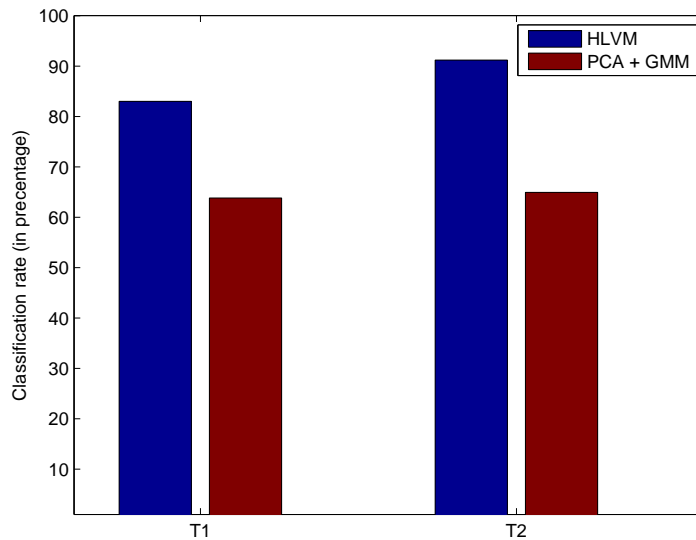


Figure 3.17: Classification results for the mouth model.

model, with sad expression being completely misclassified. This is due to the natural flexibility of the mouth which results in the highly non-linear behaviour observed here and by the way this has reduced the reliability and accuracy of captured face images during the tracking process.

3.5 Discussion

In this chapter, we have presented a foundation in the form of hierarchical decomposition, both in the conceptual and in the data sense, which offers an alternative way to represent and model facial expressions. It overcomes problems of non-linearities caused by variations in facial features and allows their description using intrinsic functionalities of the components (improved performance using eye and mouth regions was also observed by [4]). However we would like to point out that such decomposition, especially using shape alone, would not necessarily be sufficient to provide FACSs parameterisation due to the missing information needed to represent wrinkles and furrows in the skin. One of the solutions is to extend the number of modelled components to provide more detailed in-

formation. This was investigated by Cosker et al. [26] in which additional regions, including textural information, were included to extend the amount of available facial state data. On the other hand, using shape alone allows us to further reduce the complexity of the resulting model by being able to represent both of the eyes in a unified model.

The Hierarchical Latent Variable Model (HLVM) model provides additional benefits over PCA in the context of manifold representation. Besides its ability to model non-linear manifolds with a combination of local, linear subspaces, it also provides a density model, from which class conditional probabilities can be easily computed. However, because it utilises the EM algorithm, it is rather susceptible to initial starting positions during the training stage, so can be time consuming in creating an optimal or best fitting model for a given set of data, and its model.

If accurate and reliable tracking is available, then this model alone could be used to represent each of the face components reliably. Unfortunately, such an assumption is rarely valid in real world environments.

This investigation also highlights the importance of the eye region based on its very good performance, over the mouth region. This is partially due to the better tracking results in that area, but mainly due to the small number of classes being modelled and the relative rigidity of the sub-components that correspond to eyes and their brows.

In the next chapter we investigate how the root of our hierarchical model can be utilised to estimate the pose of the subject's face.

	Neutral	Anger	Smile	Grin	Open/Fear	Sad
Neutral	73.79%	4.85%	1.94%	19.42%	0%	0%
Anger	0%	100%	0%	0%	0%	0%
Smile	8.14%	0%	91.86%	0%	0%	0%
Grin	0%	0%	0%	100%	0%	0%
Open/Fear	0%	0%	0%	2.48%	97.52%	0%
Sad	12.13%	0	0	0	87.87%	0%

(a) Test Sequence **T1**

	Neutral	Anger	Smile	Grin	Open/Fear	Sad
Neutral	86.67%	0%	5.33%	6.67%	1.33%	0%
Anger	1.16%	98.84%	0%	0%	0%	0%
Smile	3.46%	0%	96.54%	0%	0%	0%
Grin	0%	0%	3.33%	96.67%	11.66%	0%
Open/Fear	0%	0%	0%	0%	100%	0%
Sad	18.75%	0%	0%	10.42%	70.83%	0%

(b) Test Sequence **T2**Table 3.7: Confusion matrix of the mouth state classification for **T1** (a) and **T2** test sequences in a model built using a person-specific dataset .

Chapter 4

Pose Estimation

Pose information is an important component of facial dynamics. As our head rarely stays still, it plays an important role in the formulation of the meaningful and realistic parameterisation of expressions [76, 45]. Pose can also implicitly encode expressions, such as agreement or disagreement, and enrich existing information by adding subtle tilts or movements. In this chapter we present a method for estimating pose based on a sparse training set that covers only a fraction of the viewsphere but is able to provide generalisation to a continuous viewsphere, as compared to most of the existing models that require dense sampling of the entire viewsphere [43, 80]. Rather than adopting an ad-hoc approach, our method maps directly into the hierarchical shape framework introduced in Chapter 3. This model can also be used in the synthesis of arbitrary views where it serves as a shape, or warp basis onto which a chosen texture can be rendered, either from a single image or from the underlying appearance model. Furthermore a prior knowledge of pose information can help in bootstrapping the Active Appearance Model (AAM) fitting process where, due to self-occlusion caused by large pose variation, parts of the texture information cannot be reliably extracted.

We begin by investigating the pose model in question, then provide experiments that demonstrate our findings.

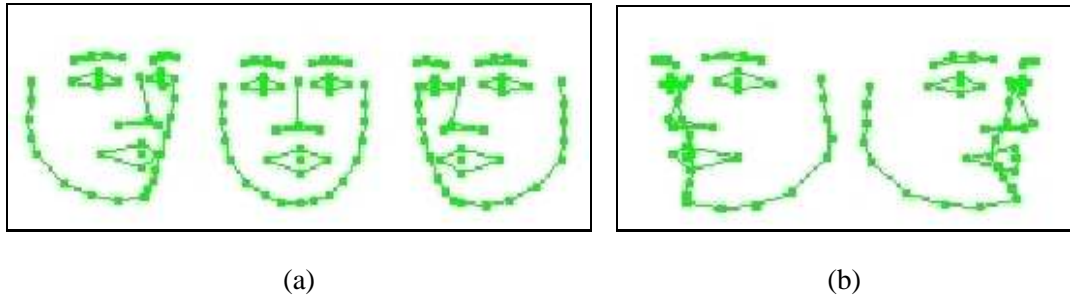


Figure 4.1: Invalid shape reconstructions of profile views (b). These were generated using modes of variations of a linear shape model trained at near frontal views (a).

4.1 Pose Model

Large pose variations cause the shape space to become highly non-linear, and this makes the linear mapping used to model the subspace no longer sufficient [43]. The problem is illustrated in Figure 4.1. Invalid shape reconstruction of profile views (b) on the right, where the underlying model was generated using samples from the near frontal views in (a). Our investigation [121] shows that the shape distribution forms distinctive bands of points in the PCA space with respect to the pose. Figure 4.2 shows projections of the data onto the first three principal axes of the shape model. The grouping was performed with respect to Y-axis (yaw) rotation, where triangles represent $[-40^\circ, -20^\circ]$, crosses $[-10^\circ, 10^\circ]$ and circles $[20^\circ, 40^\circ]$ ranges respectively, sampled at 10° intervals. We choose to model the underlying non-linear manifold with a combination of linear sub-components. The methodology is similar to that of [49] where combination of PCA models was used to constrain the Valid Shape Region (VSR). But in addition to those constraints, we also strive to provide both a pose estimation and a synthesis basis. Our pose model is based on the Hierarchical Latent Variable Model (HLVM) described in Section 3.3. We utilise the second level of the latent hierarchy which is equivalent to a mixture of PPCA models defined by Equation (3.25). Then the pose density model of the mixture is given by:

$$p(\mathbf{t}_{pse}) = \sum_{i=1}^M \pi_i p(\mathbf{t}_{pse}|i) \quad (4.1)$$

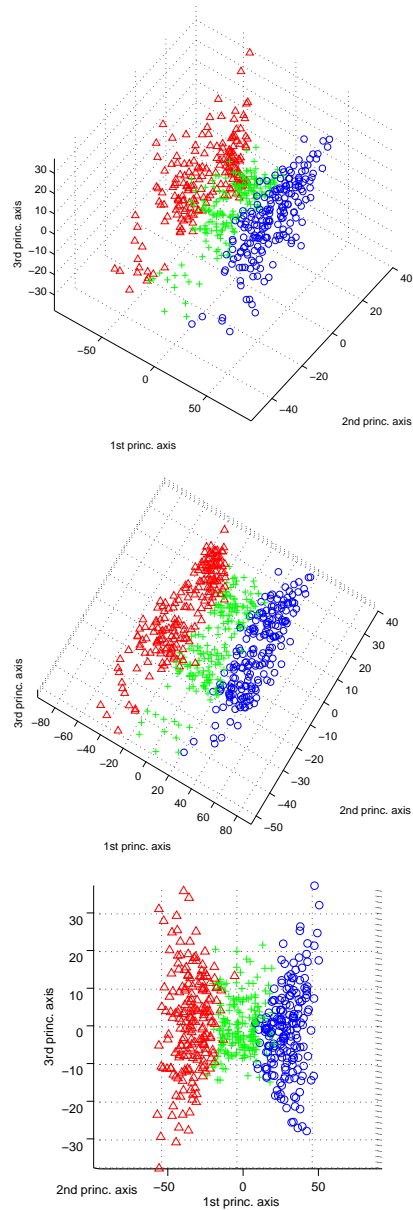


Figure 4.2: The shape variation of facial expression images from $[-40^\circ, 40^\circ]$ 3D views (in yaw) projected onto the 1st three principal components. The manifold forms continuous and separable clusters: $[-40^\circ, -20^\circ]$ (shown by triangles), $[-10^\circ, 10^\circ]$ (shown by crosses) and $[20^\circ, 40^\circ]$ (shown by circles)

where \mathbf{t}_{pse} is a pose shape vector, M defines the number of components in the mixture, and π_i are the mixing coefficients corresponding to the mixture components $p(\mathbf{t}_{pse}|i)$. A visualisation of the pose model is shown in Figure 4.3. In this figure the top row corresponds to a single PPCA model, and the second row defines the components of our mixture. Each of the mixture components corresponds to one of the rotation bands, $[-40^\circ, -20^\circ, 0^\circ, 20^\circ, 40^\circ]$, sampled at 20° intervals in yaw. We focus on yaw as the discriminative factor for our clustering scheme, as this type of rotation is the primary cause of non-linearities, and is more likely to be present in the input data (as opposed to pitch rotation). Conceptually our method is similar to the View-Based Appearance approaches of [23] but with the

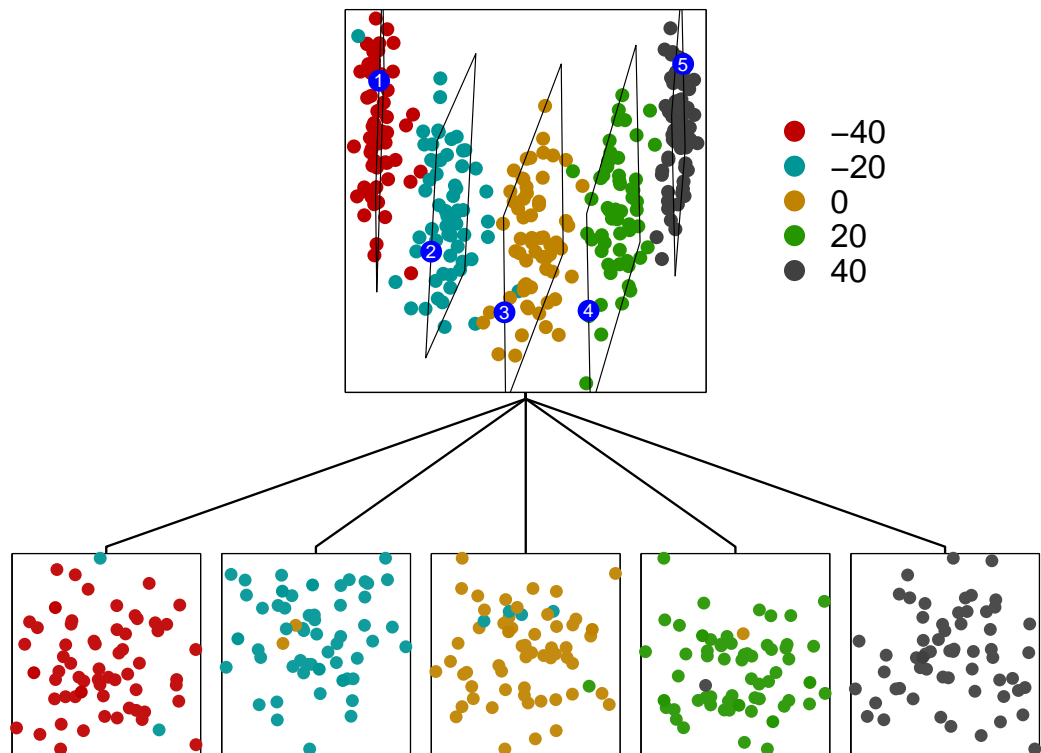


Figure 4.3: Visualisation of the hierarchical pose model. Bottom row corresponds to components of our mixture (depicted with yaw rotation sampled in 20° intervals at $[-40^\circ, -20^\circ, 0^\circ, 20^\circ, 40^\circ]$).

following differences: Firstly we do not require the manual process of building multiple models. As we use shape only, the dimensionality of the model and resulting training time, is greatly reduced.

Secondly there is no need to establish the relationship between each of the models to find which one is responsible for generating a given data point. Finally the full correspondence between features among the components is established, allowing for uniform trajectories of the shape across the model view-sphere. Our model is able to easily generate novel samples at arbitrary view-points and warp selected textures onto them. This is in contrast to [43], where KPCA was used to model resulting non-linearities and the feature space was implicit and unknown, the model proposed here can easily reconstruct, or generate novel samples, as the model defines both forward and reverse mappings to and from lower dimensional manifold. We chose not to incorporate any temporal information, and the estimation can be done frame-wise on-the-fly in real time. This approach is able to cope with very large jumps and discontinuities in pose change.

Each of the components is capable of capturing an underlying variation of interest, according to its clustering scheme. Figure 4.4 shows a visualisation of the mean values corresponding to each of the PPCA components with respective yaw labels determined by the predefined data grouping.

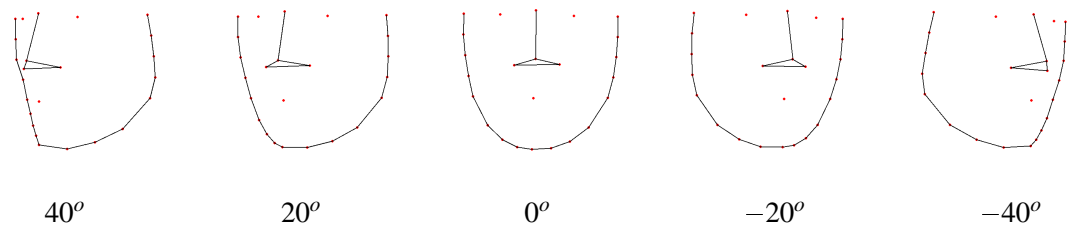


Figure 4.4: Means corresponding to each of the PPCA models defined by the hierarchical pose model with their respective yaw labels.

From the definition of our model, latent variables are assumed to be drawn from a Gaussian distribution. Given any d -dimensional multivariate Gaussian distribution with mean μ and covariance matrix \mathbf{C} , its marginal q -dimensional multivariate distribution (where $q \ll d$) is also Gaussian, as described by Krzanowski [62]. Let \mathbf{B} be a $q \times d$ dimensional matrix with diagonal elements set to 1 and all remaining elements equal to 0. Then the marginal q -component multivariate Probability

Distribution Function (PDF) f_q is given by:

$$f_q \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu}, \mathbf{BCB}^T) \quad (4.2)$$

The corresponding Cumulative Distribution Function (CDF) is defined as the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x . If we interpret this as a value less than or equal to a specific pose value captured by the underlying distribution, we can use this to extract the pose information based on it. Following the concept of the marginal PDF, we define the CDF Φ such that for a q -dimensional random variable \mathbf{x} it is given by:

$$\Phi(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_q(\mathbf{x}) d\mathbf{x} \quad (4.3)$$

We are mostly interested in CDFs that are closely related to the components responsible for yaw and pitch rotation. For a given shape \mathbf{t}_{pse} the estimate of the pitch rotation r_{pitch} is given by Equation (4.4), where a_1, a_2 are coefficients of a linear polynomial and p_{pitch} is the marginal CDF of the model component responsible for generating shape \mathbf{t}_{pse} :

$$\begin{aligned} r_{pitch} &= a_1 p_{pitch} + a_2 \\ p_{pitch} &= \Phi_{mp}(\mathbf{t}_{pse}) \end{aligned} \quad (4.4)$$

Figure 4.5 shows visualisation of linear relationship for the pitch estimate for an example cluster.

For yaw rotation r_{yaw} the estimate is given by Equation (4.5), where b_1, b_2 are coefficients of a linear polynomial and p_{yaw} is a weighted sum of marginal CDFs of the model.

$$\begin{aligned} r_{yaw} &= b_1 p_{yaw} + b_2 \\ p_{yaw} &= \sum_{i=1}^M \pi_i \Phi_{my}(\mathbf{t}_{pse}|i) \end{aligned} \quad (4.5)$$

Figure 4.6 shows visualisation of linear relationship for the yaw estimate.

Although we do not account for roll rotation in our model, we adopted the approach of Horprasert et al. [52], where for the eye centroids \mathbf{t}_{eyeL} and \mathbf{t}_{eyeR} , the head roll is given by:

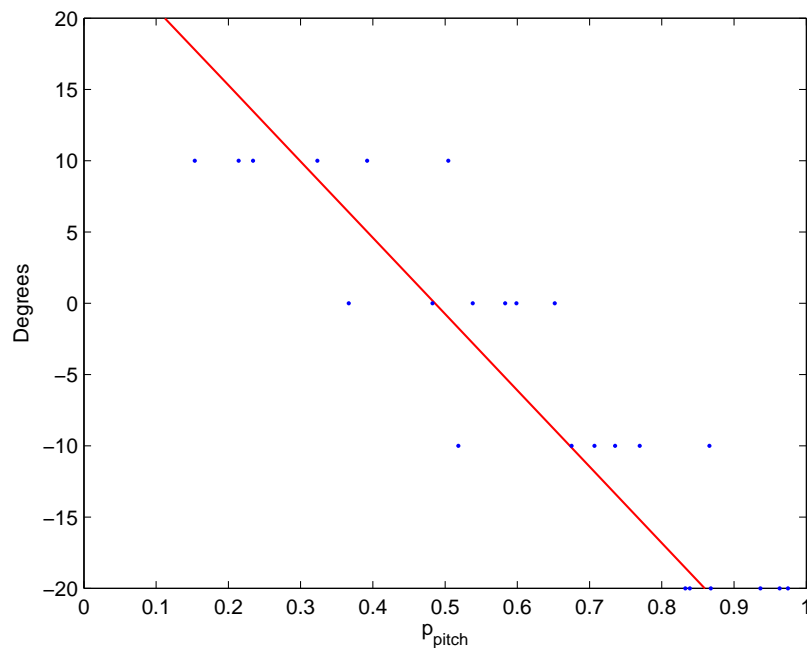


Figure 4.5: Visualisation of linear relationship for the pitch estimate for an example cluster using discretely sampled training data at 10° intervals.

$$\begin{aligned}
 r_{roll} &= \arctan \frac{\Delta y}{\Delta x} \\
 \Delta y &= \mathbf{t}_{eyeR}^y - \mathbf{t}_{eyeL}^y \\
 \Delta x &= \mathbf{t}_{eyeR}^x - \mathbf{t}_{eyeL}^x
 \end{aligned} \tag{4.6}$$

The above equations encompass the formulation of our model.

4.2 Experiment

Our training and validation dataset consists of 540 labelled 2D shapes from 12 individuals each defining 14 landmarks. Figure 4.7 shows selected training samples from this training set. The shapes have associated ground truth information, obtained by using a magnetic sensor rigidly attached to the subjects head, and are sparsely sampled at 10° intervals, covering only part of the view-sphere, $(-40^\circ, 40^\circ)$ around the yaw axis and $(-20^\circ, 20^\circ)$ around the pitch axis. We divided the dataset

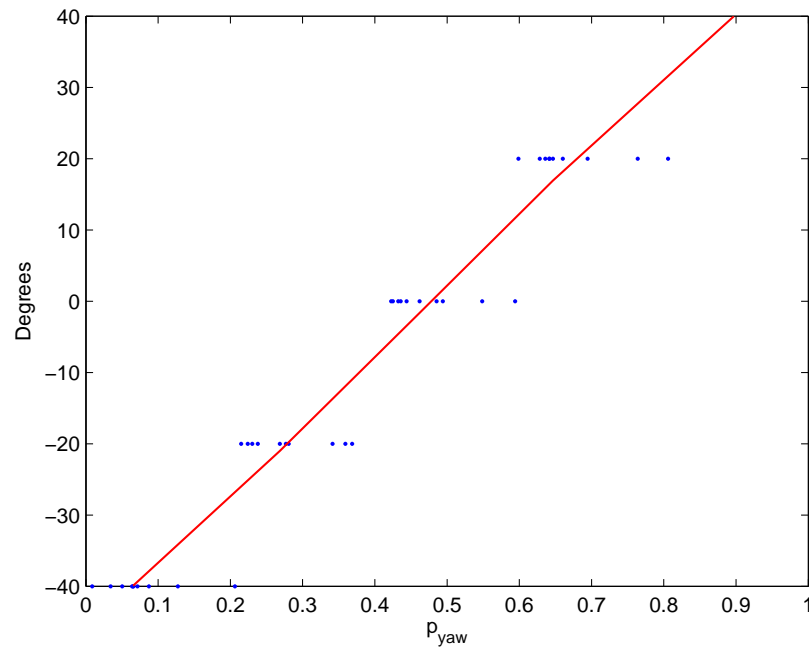


Figure 4.6: Visualisation of the linear relationship for the yaw estimate using discretely sampled training data at 20° intervals.



Figure 4.7: Selected training samples from the pose estimator dataset.

into groups consisting of 300 and 240 samples. The first group contained samples selected using 20° intervals for yaw, consisted of rotation bands $[-40^\circ, -20^\circ, 0^\circ, 20^\circ, 40^\circ]$, and was used for training the model. The second group was set aside for validation purposes, sampled using similar 20° intervals and consisted of yaw rotation bands not used in the training set, i.e. $[-30^\circ, -10^\circ, 10^\circ, 30^\circ]$.

To train the model we chose to calculate the means for each of the groups of labelled data, and

used those means as starting points for the EM algorithm. Although we experimented with using the K-means algorithm we found that the expected clustering arrangement, with respect to location of clusters, was not always guaranteed. And as our method was based on the assumption of a pre-defined clustering scheme we settled for the former EM approach. Following mean calculation, the EM algorithm was run and after a few iterations convergence was declared. In contrast to the approach that was undertaken when building the models for the eye and mouth components in Section 3.4, this offers a fully automated model building process without the need for user-driven feedback. For each of the individuals we have calculated corresponding marginal CDFs and fitted appropriate polynomials according to Equation (4.5) and Equation (4.6) for pitch and yaw respectively. Also, because there were differences in the estimated parameters, final polynomial coefficients were taken as an average over the coefficients obtained from several different individuals.

Facial expression can affect the resulting pose information obtained by the model. This is mainly caused by variations, or movement of facial components (predominantly the eyes and mouth in the case of our shape representation). Although techniques try to model the pose by including these regions and account for their influence, we choose the alternative approach of selecting only those features that will remove, or minimise the impact caused by varying facial expression. We choose jaw and nose outlines, and the centres of the eyes and mouth for our pose shape representation. Our model is able to generalise to a denser shape model. This is achieved by down-sampling the larger PDM to the required size by calculating the centroids of the eyes and mouth and selecting a subset of jaw outline landmarks. For the mouth shape, the centroid is calculated as an intersection of two lines formed by two middle landmarks on the upper and lower lips and two landmarks on the jaw outline. This provides a more stable representation under varying facial expression and compensates for more flexible and unconstrained lip deformations. Figure 4.8 shows how eye and mouth centroids are determined.

Figure 4.9 provides visualisation of this down-sampling process across selected facial expressions using previously unseen data. The top row corresponds to the full PDM representation consisting of 74 landmarks and the bottom row to the down-sampled 14 landmark representation. Each of the



Figure 4.8: Visualisation of the eye and mouth centroids with the latter calculated by intersecting two lines formed by the landmarks on lip and jaw outline.

down-sampled shapes has a pose estimate associated with it (yaw,pitch,roll), and we can see that there is not much pose change between different expressions when using the reduced representation, compared to that exhibited in the full size shape model. We acknowledge that this is only an approximation, as the calculated eye and mouth centroids will exhibit small amounts of movement where facial expressions are present. However as we are more likely to experience those expressions in a frontal, or near frontal views, the impact on the estimation at extreme views is minimised.

To test the ability of the model to generalise to new data we employ the second part of the dataset. We project the data onto the latent subspace, estimate its pose information and compare it against the assigned ground truth labels. Given the Root Mean Square Error (RMSE) as our measure of error:

$$RMSE(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \mathbf{y}_i)^2} \quad (4.7)$$

where \mathbf{x} represents the vector of ground truth values and \mathbf{y} is corresponding vector of estimated values, we obtained average errors of 5.22 degrees for yaw and 6.66 degrees for pitch.

To perform an evaluation on how well the model is able to generalise to a continuous sequence we have used a test set of 500 samples which exhibited pose variations in yaw, pitch and roll with

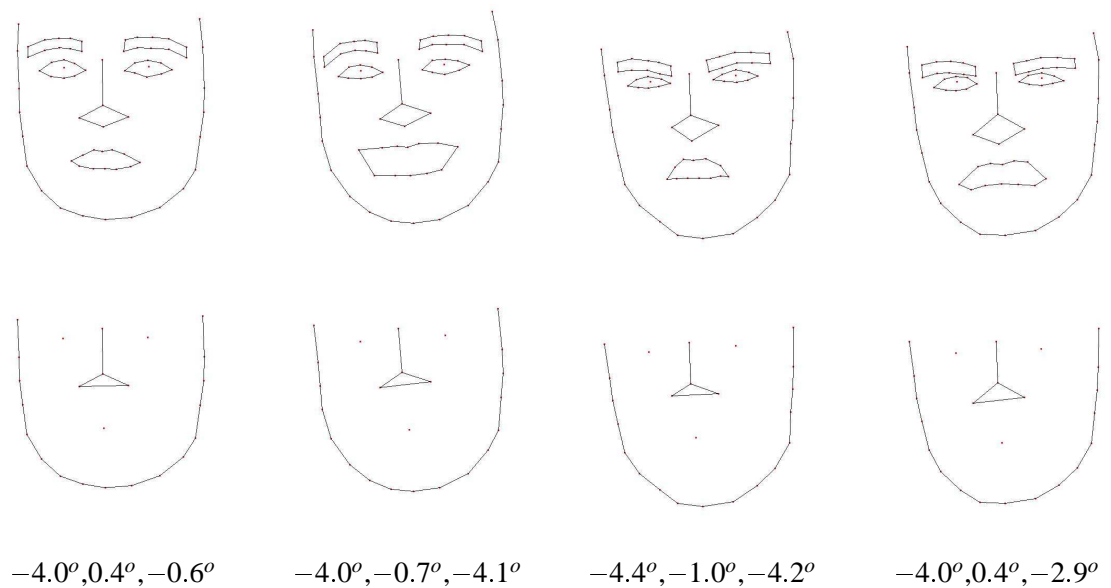


Figure 4.9: Denser shape model consisting of 74 landmarks (top), sparse set of 14 landmarks (bottom) with corresponding estimated pose angles (yaw,pitch,roll).

a much denser arrangement of landmark points. Each of the shapes was down-sampled and subsequently used to estimate the pose. Due to the lack of a tracker that can reliably and consistently track a face beyond frontal, or near frontal views we have manually labelled the training set. The top plot in Figure 4.10 shows the result of the experiment. We can see that the yaw estimation is consistent throughout the sequence, but pitch estimation contains errors, especially around frames 228 – 285 and 340 – 385. This is caused by an incorrect choice of cluster membership obtained through class-conditional probability. The errors show that our initial assumption that each of the clusters responsible for the appropriate yaw rotation band will always be chosen to estimate the pitch was not correct, and it also highlights the fact that the clusters outside the designated band perform poorly in such estimation. To address this problem we have introduced yaw-based constraints, which follow the design of the model more closely. Rather than using class-conditional probability to determine the cluster membership, we enforce the selection by conditioning each choice on the interval into which the estimated yaw value falls. The bottom plot in Figure 4.10 demonstrates the results

obtained from this yaw constrained model.

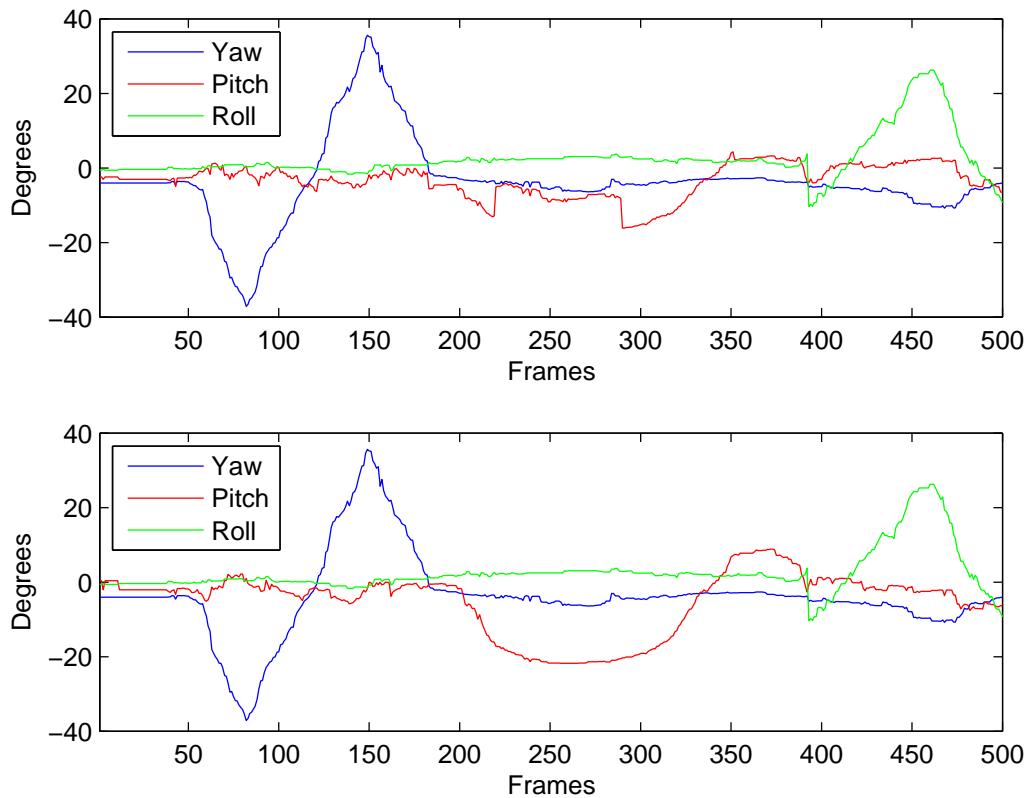


Figure 4.10: Pose estimates obtained from the original pose model (top) and from the yaw constrained one (bottom).

Although introducing the constraints corrected the previous errors, both methods produce jagged curves caused by small shifts in landmarks, and general inaccuracies resulting from the estimation process. This makes the representation unsuitable for parameterisation, as they will produce jerky animation sequences. In order to produce smoother curves we convolved the estimated pose data with a Gaussian kernel. Although some of the information is lost during the smoothing process, this is a justified trade-off and should not be too noticeable to the viewer in most circumstances. Figure 4.11 shows the results of this filtering, where the difference between the original (jagged) curve and the smoothed one can clearly be seen. The final results are depicted in Figure 4.12 with

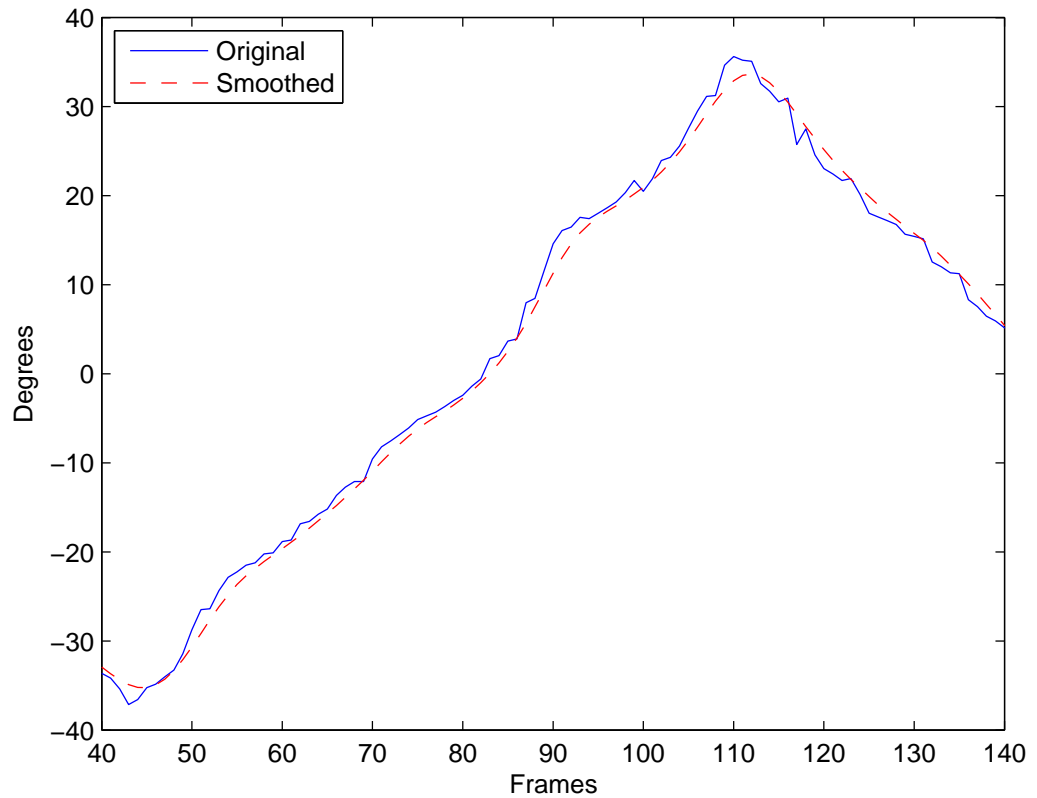


Figure 4.11: Section of pitch rotation plot: original jagged curves (green solid line) and their smoothed out counterparts (red dashed line).

the yaw corrected and smoothed rotation curves. Figure 4.13 shows the selected frames from the input sequence (left image), with the corresponding synthesised avatar alongside (right image). The synthesis was performed with LightWave 3D¹, where each of the rotation curves was converted into its respective channel envelope and used to drive the avatar. The resulting video can be viewed here².

The experiments conducted in this section have shown how the pose model developed here can be used to provide smooth and continuous pose estimation.

¹LightWave 3D <http://www.newtek.com/lightwave/>

²http://www.eecs.qmul.ac.uk/~lukas/videos/pose_*.avi

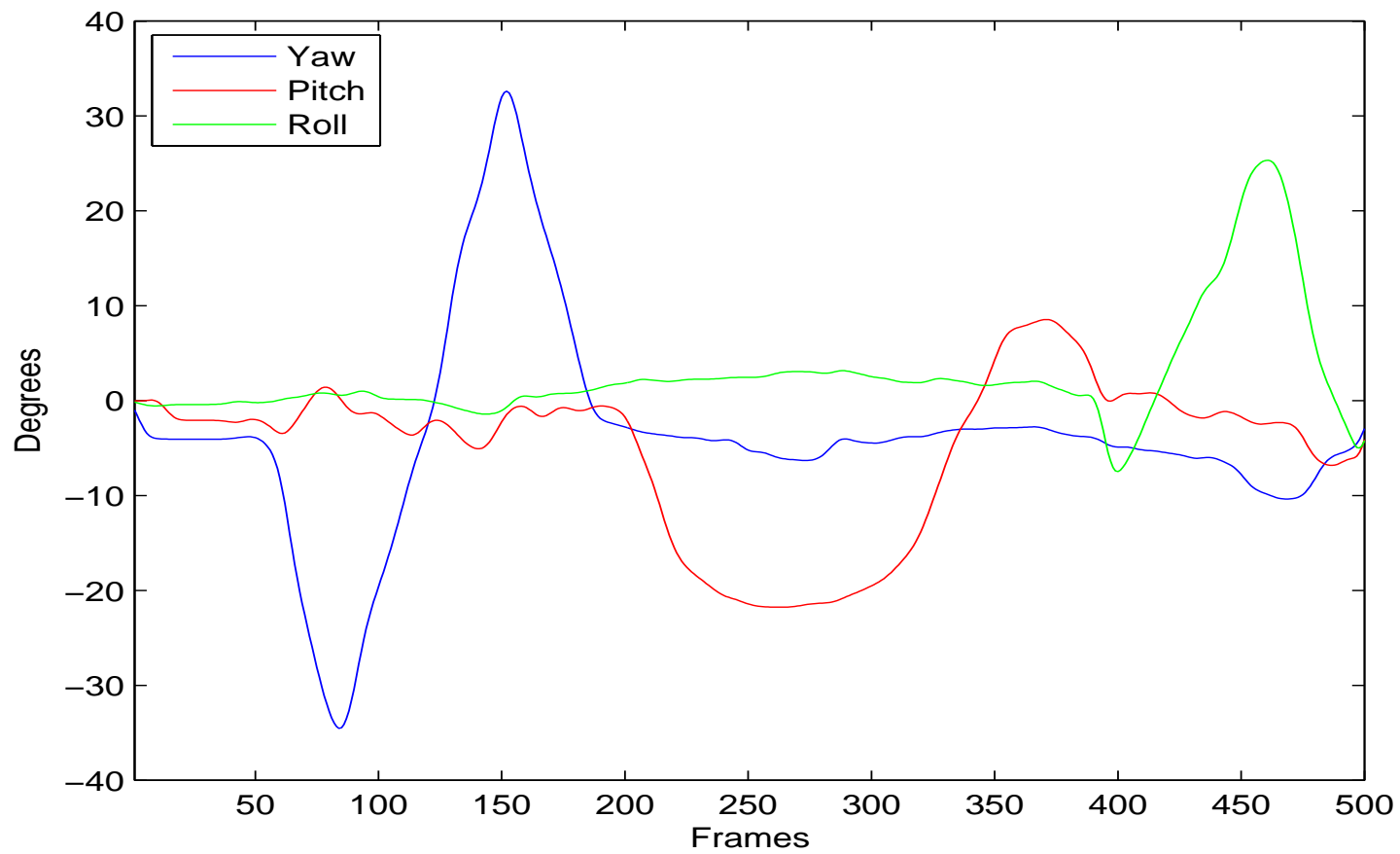


Figure 4.12: Results of continuous pose estimation experiment.

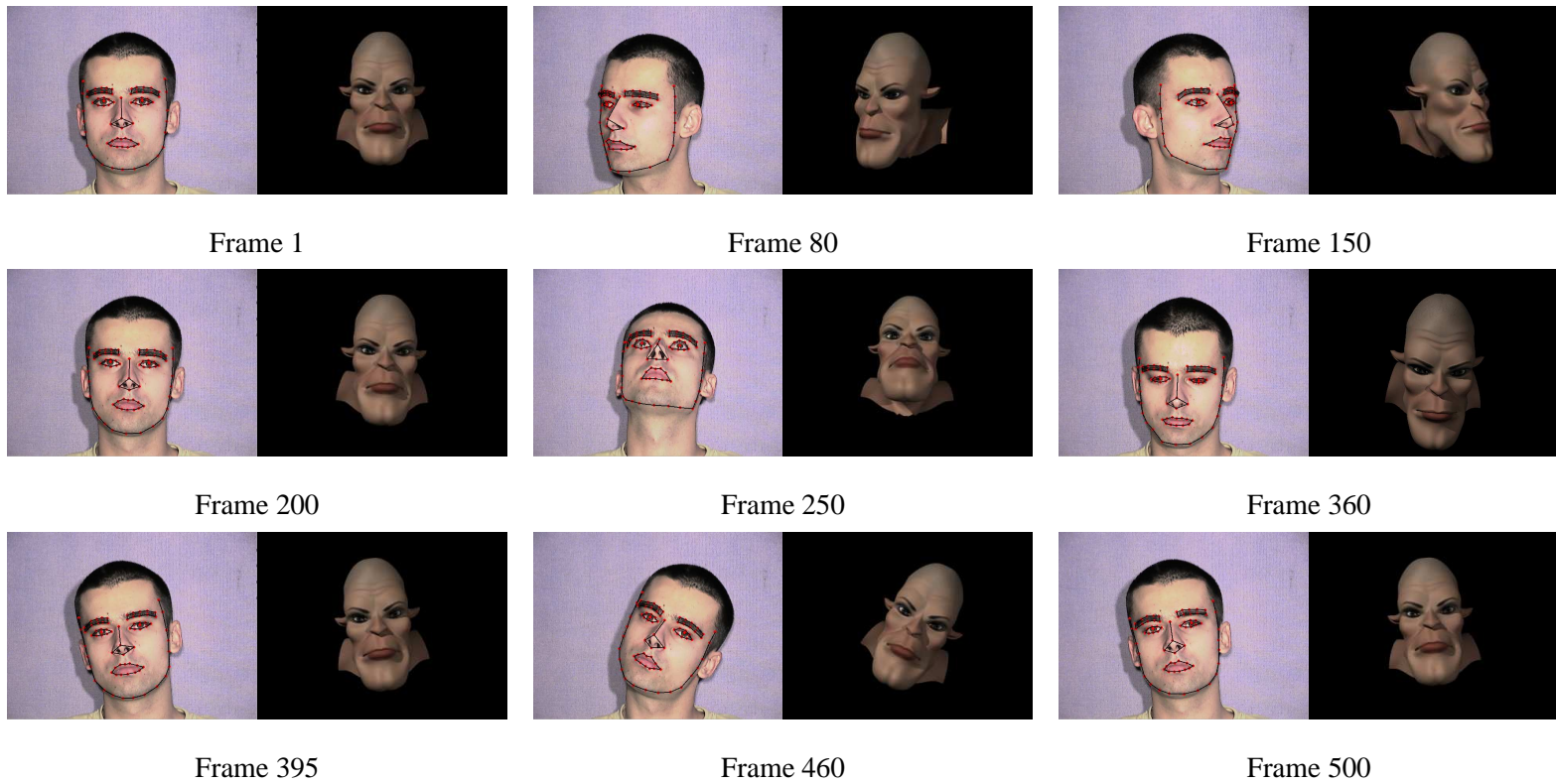


Figure 4.13: Selected samples from the original sequence (left image) with corresponding frames from the synthesised avatar (right image).

4.3 Synthesis

The pose model developed in this chapter can be also be used to synthesise samples at arbitrary viewpoints. We investigate the use of Independent Appearance Models (IAMs), where dependence, or correspondence, between texture and shape is ignored. Although this increases the overall dimensionality of the model, it opens up the possibility of using different texture models interchangeably with a single shape model. In our case the given shape is treated as a warp, or morph basis onto which a selected texture will be rendered from the underlying appearance model.

This procedure usually involves a shape-free texture \mathbf{t} which will be warped onto a target shape \mathbf{s}_t , and using shape \mathbf{t}_b as the warp base, which is usually defined by the mean of the model. For each of the shapes we will have a corresponding set of triangles that have been obtained obtaining by applying Delaunay triangulation [64] to the convex hull defined by the set of control points, or landmarks. The number of triangles, and their respective vertex indices are computed once, during the shape model creation. This is to reduce the computational load caused by re-calculating the triangulation every time warping is performed, but more importantly to ensure that the rendering of the texture stays uniform across the synthesis and extraction process. The triangulation can be done in an automated way, where the triangles are created to best fit the control points, or they can be manually defined to conform to a particular scheme, such as from left to right, or from top to bottom.

Next, for each of the triangles, the shape-free texture fragment \mathbf{t} , contained within, is warped from the base shape \mathbf{t}_b onto the target shape \mathbf{t}_t to produce the resulting synthesised sample. This process works well in frontal, or near frontal views, but fails in extreme views when a linear model has been used to generate the view. Figure 4.14 demonstrates such failure: (a) shows shape-free base, or mean shape, with its corresponding triangulation. (b) and (c) show the shape reconstructions of extreme views and their corresponding triangulation from a linear shape model trained at near frontal views. Because of the inability of the model to generate valid shapes at these views, some of the triangles will overlap. During the warping process this will cause distortion to the resulting texture, at arbitrary places if the triangulation was automated, or at the edges of a single extreme pose view if the triangulation was performed left to right.

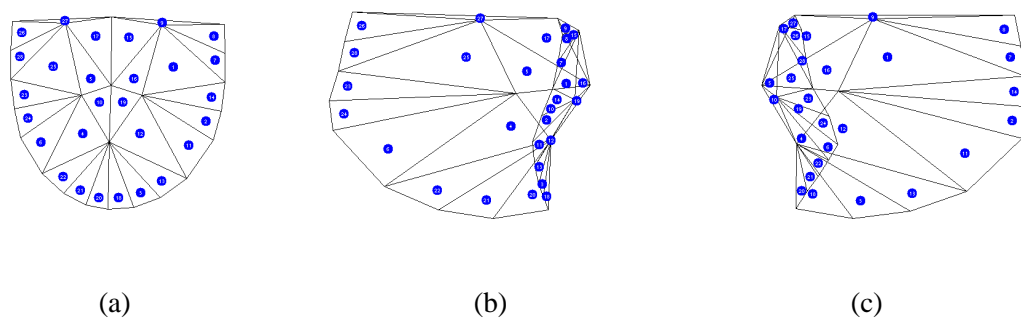


Figure 4.14: Visualisation of triangulation when warping using a linear model: (a) base (mean) triangulation, (b, c) extreme view triangulation.

Our hierarchical model is able to handle the non-linear changes much better. Figure 4.15 shows

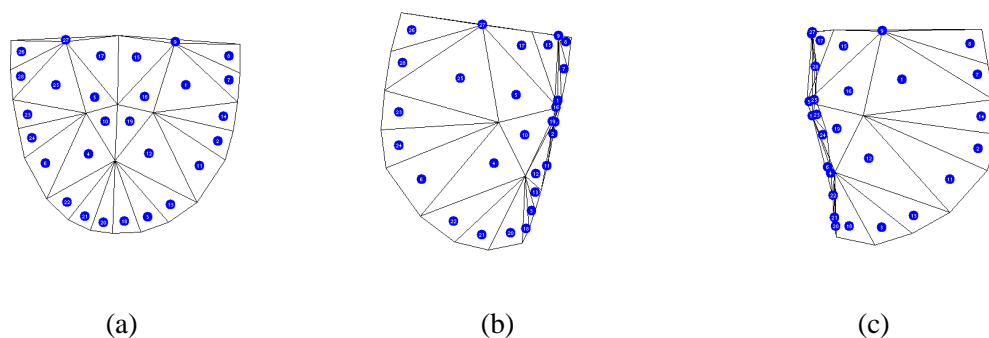


Figure 4.15: Visualisation of triangulation when warping using a hierarchical model: (a) base (mean) triangulation, (b, c) extreme view triangulation.

a similar triangulation where (a) is the shape-free base, or mean shape, and (b) and (c) are the shape reconstructions of profile views.

Our choice of features for the pose shape model was driven by the need to minimise, or remove, the effects of facial expressions. Interestingly, in the context of synthesis, or warping this has an inverse effect. Rather than having shape free appearance which will minimise the effects of the

expressions on the underlying appearance model, we have an appearance model which does not aim to minimise the effects of the expressions. This is illustrated in Figure 4.16, where the top row corresponds to a full size PDM with a small amount of expression being carried over to the resulting shape free appearance sample. The bottom row depicts the pose PDM where most of the expression

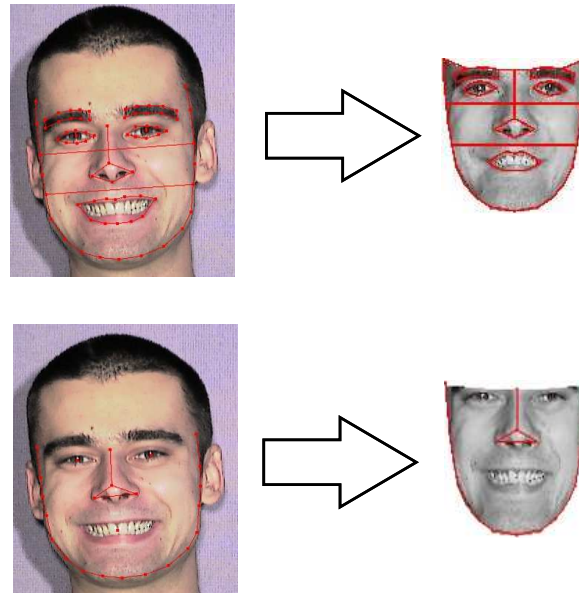
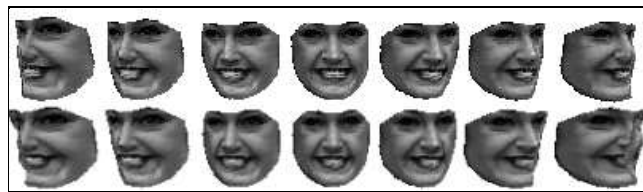


Figure 4.16: Shape and expression-free appearance with full PDM mask (top) and shape-free with pose PDM mask (bottom).

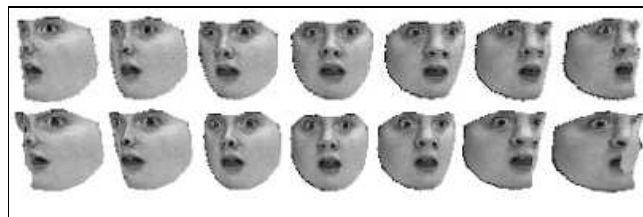
is propagated over to the corresponding shape free appearance sample. Given a separate texture model, and our pose model, we can easily generate arbitrary expressions at arbitrary viewpoints by varying the components of each of the models independently. To show this we built an appearance model using 490 images from the Cohn-Kanade database, and used it as a basis for synthesis of arbitrary viewpoints. Figure 4.17 shows samples of the results from our synthesis, where in each set of expressions, labelled (a), (b) and (c), the top row corresponds to the expression synthesised using our pose model and bottom row to an expression synthesised using a PCA pose model, and Figure 4.18 illustrates the cluster membership for the shape bases used in the synthesis. Note that in the PCA model synthesis, only one of the extreme views exhibits severe texture distortion. This is due to the triangulation order that we imposed (left to right), so that overlapping triangles happen



(a) Expression 1



(b) Expression 2



(c) Expression 3

Figure 4.17: Examples of morphing (synthesising) facial expressions into extreme virtual views. The top rows in (a), (b), (c) were computed using the HLVM model. The bottom rows were computed using the PCA model with visible kinks at extreme 3D views (profile views) due to the non-linearities present.

to be drawn in a correct order (as an analogy, this can also be thought of as a painter's algorithm [113]). If the pose information were known, the warping process could be controlled by enforcing the order with respect to the pose value, i.e. either from left to right or from right to left. This would ensure that the incorrect triangles are drawn first, and correct ones second. However this would be a half measure and in the case of the linear model this would only work for overlapping triangles and would not correct the artifacts caused by the extreme, non-overlapping distortions. In this section we

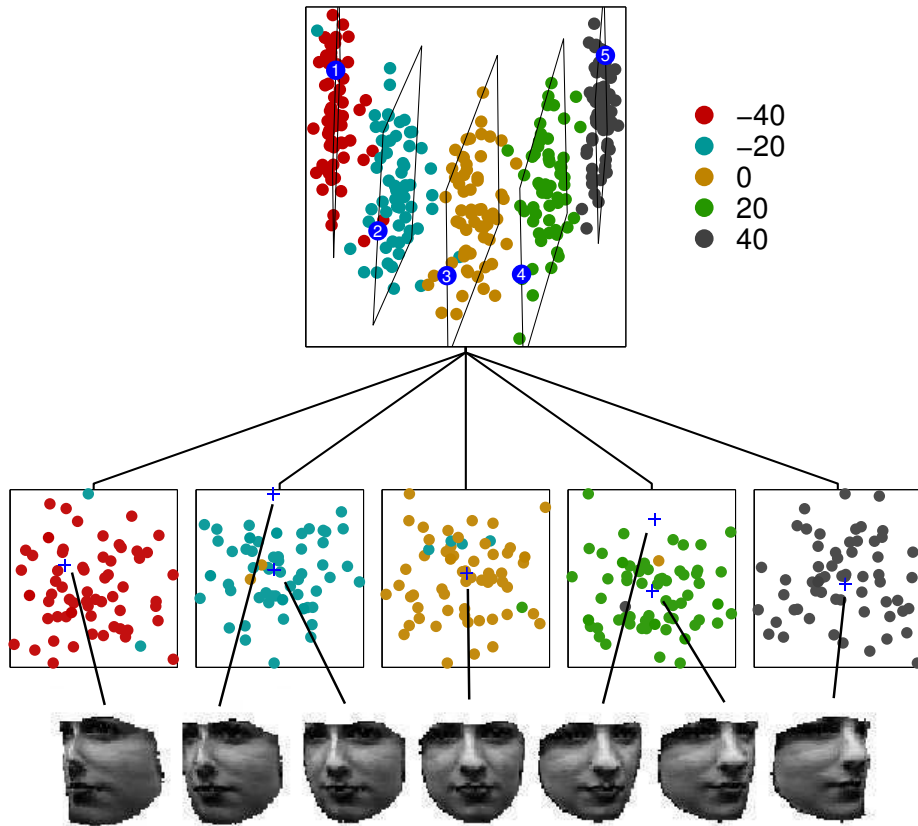


Figure 4.18: Visualisation of the cluster membership for the shape bases used in the synthesis (top and middle rows) together with the synthesised samples (bottom row).

have described and demonstrated how the pose model can be used to successfully generate synthetic samples at arbitrary viewpoints.

4.4 Improved AAM Fitting

The fitting process of the AAM is defined as an optimisation problem, where the difference between an image and the synthesised counterpart is minimised. A difference vector can be defined by:

$$\Delta\mathbf{T} = \mathbf{T}_{im} - \mathbf{T}_m \quad (4.8)$$

where \mathbf{T}_{im} , \mathbf{T}_m are the texture instances in the image and model frame respectively. However the basic representation of an AAM is only able to cope with frontal or near-frontal views ($[-15^\circ, 15^\circ]$)

in yaw). At the extremes of pose change, due to occlusion during the warping process, the texture is distorted creating large residuals and causing tracking failure. Figure 4.19 shows the original images from a sequence in the top row, and the corresponding frontal view warped texture vectors, with visible distortions in the bottom row. Given the knowledge of the pose information Dornaika and

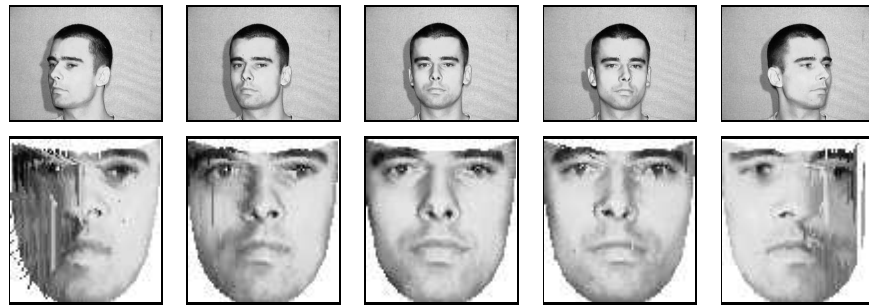


Figure 4.19: Distortions due to the pose changes and self-occlusion. Top row: original images, bottom row: frontal view warped images.

Ahlberg [29] proposed the approximation of the missing information by mirroring the warped image when necessary. Figure 4.20 shows the results of this mirroring process, where the top row consists of the original images and the bottom row shows pose corrected mirrored images. As we can see

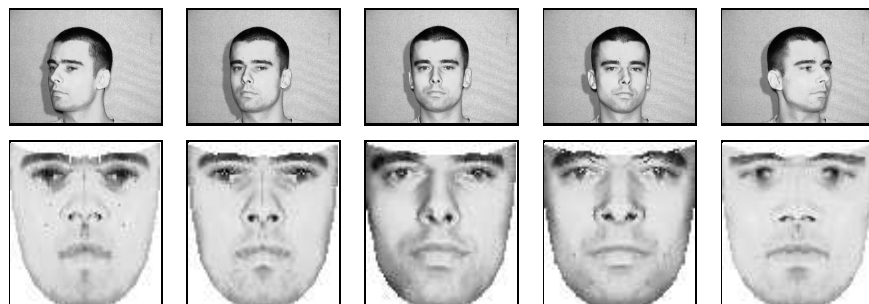


Figure 4.20: The mirroring process. Original images (top row) and resulting pose corrected frontal view warped images (bottom row).

mirroring provides only an approximation to the true representation of the face at extreme views and

can distort the true representation of the data.

To further improve the tracking process we introduce a pose corrected weight vector such that the original texture difference $\Delta\mathbf{T} = \mathbf{T}_{im} - \mathbf{T}_m$ becomes $\Delta\mathbf{T}_{corr} = \mathbf{W} \otimes \Delta\mathbf{T}$, where \mathbf{W} is the pose dependent weight vector drawn from the normal distribution and \otimes is element-wise multiplication. We consider that the weight vector as a measure of the influence of each of the pixels within the texture vector towards the final texture difference conditioned on the current pose estimate. For areas that will be occluded, hence containing distorted data, the amount of influence will be minimal, or zero and should not influence the fitting process. If we define a $w \times h$ shape free mask, then for every row of pixels in that mask, each element j in that row contains the weight value given by its probability:

$$p(j) = e^{-\frac{(j-\mu)^2}{2\sigma^2}} \quad (4.9)$$

$$\mu = 0.5 * w \quad (4.10)$$

$$\sigma = \frac{0.5 * E * (0.5 * w)}{90} \quad (4.11)$$

where $E = |40 - r_{yaw}|$ when $|r_{yaw}| < 40$ or $E = 1$ otherwise. We only calculate the weights for the occluded half of the image, which we determine from the pose estimation, and the fully visible parts of the image are set to 1. Figure 4.21 shows different representations of \mathbf{W} with respect to different yaw rotation values.

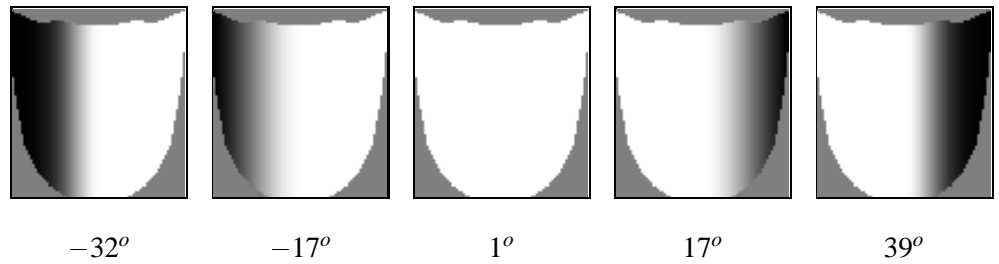


Figure 4.21: Different weight vector representations for different yaw rotation values.

We have compared the performance of the original formulation of the AAM with our method.

Figure 4.22 shows the results of the experiment, where the top row contains the frames from the original AAM implementation, and the bottom row contains the pose corrected method. Although we obtained better results, our approach was not able to provide consistent and reliable tracking. This is due to the influence of the underlying linear shape model, and its flexible and invalid variation.

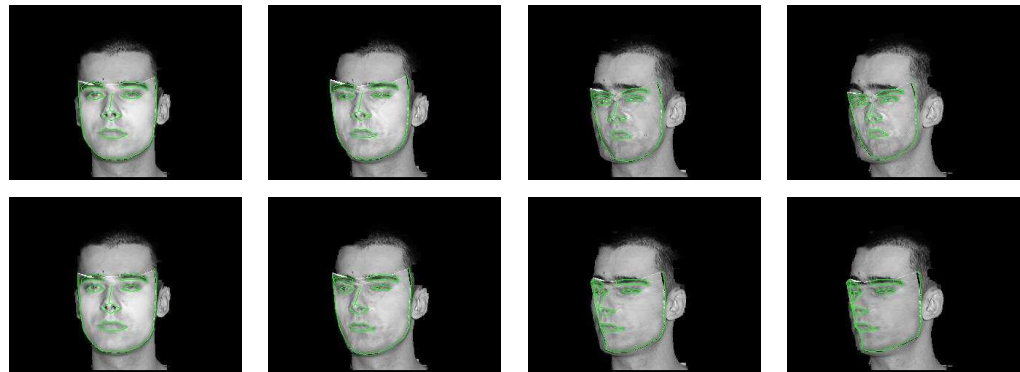


Figure 4.22: Selected frames from the experiment in fitting AAM onto the extreme pose view. Top row corresponds to the original AAM model formulation and bottom row to the pose corrected model.

In summary, we have investigated and shown how to minimise the effects of distortions due to self-occlusions caused by extreme pose viewpoints. The results show improvement over standard AAM fitting techniques in the literature.

4.5 Discussion

In this chapter we have described a pose model, which integrates into our hierarchical decomposition, and is able to estimate and generalise to continuous poses based on a sparse set of training samples. The choice of shape component and the underlying features, allowed us to minimise the influence that facial expressions have on the resulting pose estimation, a goal that would be difficult to achieve if the appearance component was considered. However in some instances, for example in the case of the surprise expression, jaw outline may exhibit some amount of movement, which can affect the estimated pose. Given the underlying probabilistic framework and its ability to deal with missing

data, features such as parts of the jaw outline and centroids of eyes and the mouth can be treated as missing, and the pose can be estimated in such a way that almost completely negates the effects of expressions. Luthi et al. [74] used the missing data approach to reconstruct a full set of data from a single PPCA model using only a subset of the original features.

Our model can also be utilised as a warp basis, onto which arbitrary textures are rendered, and which provides superior performance to the linear PCA model. However our model is a combination of linear models, and invalid reconstructions can still occur in some circumstances. Extending the model to cover additional yaw rotation would extend the model coverage. Although we only consider single images from the underlying model, this could easily be extended to sequences containing the expressions exhibiting facial actions. Then the trajectories in texture space could be learned and reconstructed using an approach similar to [12].

Finally prior knowledge of the pose can help bootstrap the AAM fitting process where self-occlusions lead to missing information, and artifacts in the resulting shape free texture, affecting calculation of the pixel difference, and resulting parameter updates. Although the proposals of Section 4.4 method offered improvement over the standard AAM approach, it seems that constraining texture alone is not enough. In order to achieve stable and consistent tracking we must add constraints to our linear model, or we must adopt a non-linear model for the shape component.

Next in Chapter 5 we explore the concept of fusion, which combines the information extracted from our hierarchical components to produce final expression labels, and allows us to estimate the underlying expression intensity. Labels together with pose and intensity information are converted to a parametric form which will be used to animate a synthetic head avatar.

Chapter 5

Modelling Expression Dynamics

Facial expressions play an important role in communicating human emotions. Unfortunately the majority of approaches focus predominantly on the ability to just determine a set of emotional labels. Although frequently underestimated, the intensity or severity of these states, combined with pose information, is an important factor that contributes to the underlying dynamics of those states. Since our eyes are able to distinguish even the slightest of imperfections and because we perceive expressions mostly in a dynamic context, we should make use of this valuable information to more fully accomplish its successful transfer to a synthetic counterpart.

In this chapter we introduce a facial expression modelling framework which produces parameterised expression dynamics information. Based on our hierarchical decomposition, presented in Chapter 3, we explore the use of rule based classifiers to combine the information obtained from the facial expression components previously considered. To complete the parameterisation, we compute severity information for each of the identified expressions and produce continuous animation curves that can be used to animate a synthetic head avatar in a morph-based fashion. We investigate the use of static information only and its sufficiency for the task at hand.

Using an Active Appearance Model (AAM) based tracker, we compare the performance of our

approach with that of a holistic representation plus Bayesian Network (BN) approach under challenging conditions, where misalignments during the tracking process can produce noisy or incorrect information.

The work in this chapter begins with a description of expression parameterisation, then we explain our 3D animation model and provides experiments that demonstrate how our system performs dynamic expression modelling.

5.1 Expression Parameterisation

Let us first develop a compact parametric description of the expressions which can summarise their underlying dynamics and can be subsequently used to animate an avatar. In the context of parameterisation much recent work has been focused on recognition of Facial Action Coding System (FACS) Action Units (AUs) [67, 86, 110, 73]. We try to avoid the complexity and computational demands of processing FACS, although their detailed description of expressions and realistic animation provide a desirable goal. In real world scenarios this level of detailed information is unlikely to be easy to extract or accurate enough.

In theory each of the hierarchical components introduced in Chapter 3 should be sufficient on their own to define the resulting labels and the underlying dynamics. However in practice due to the inevitably inaccurate input information this approach is likely to produce biased results. Conceptually we wish to represent final expression labels and their underlying dynamics as a combination of intrinsic functionalities, or states of the hierarchical subcomponents. This can be summarised as:

$$expression = state_{mouth} + state_{eyeR} + state_{eyeL}$$

This concept is analogous to the way character rigging is performed in facial animation, where actions of the most salient, or influential components are combined together to produce the final state [42]. To achieve that goal, we investigate a fusion approach. Fusion is usually associated with combining data from multi-sensory input sources, but in our case we treat each of our hierarchical regions as a different input source.

5.1.1 Framework Overview

For a given image sequence, the flow of information through our processing pipeline can be described by the following tasks:

1. **Pre-processing:** The input image is converted to HSV space, which has proved to be successful in detecting, or separating, skin regions [54]. Background segmentation is performed by thresholding Hue and Saturation components to segment the image into the face region (foreground) and anything else (background). Next the segmented image is converted to grey scale and passed on to the subsequent module.
2. **Tracking:** The face of interest is tracked across the sequence of images. Once the image has been fitted, extraction of the shape component and its hierarchical de-composition is performed.
3. **Pose estimation:** The root component of the hierarchical decomposition is used to estimate yaw, pitch and roll rotations. The information can be fed back to the tracker to assist with non-frontal views.
4. **Label assignment:** Each of the hierarchical components has its final expression label determined.
5. **Intensity estimation:** Once the labels have been assigned the corresponding intensity for that label is calculated.
6. **Parameterisation:** Pose information, together with labels and corresponding intensity is converted into a parametric description of the facial state.
7. **Rendering:** The resulting expression is rendered using the previously defined parametric form.

The overview of our framework is depicted in Figure 5.1. The bottom line of boxes corresponds to the framework described. The upper boxes in Figure 5.1 correspond to training the expression model and training the pose model.

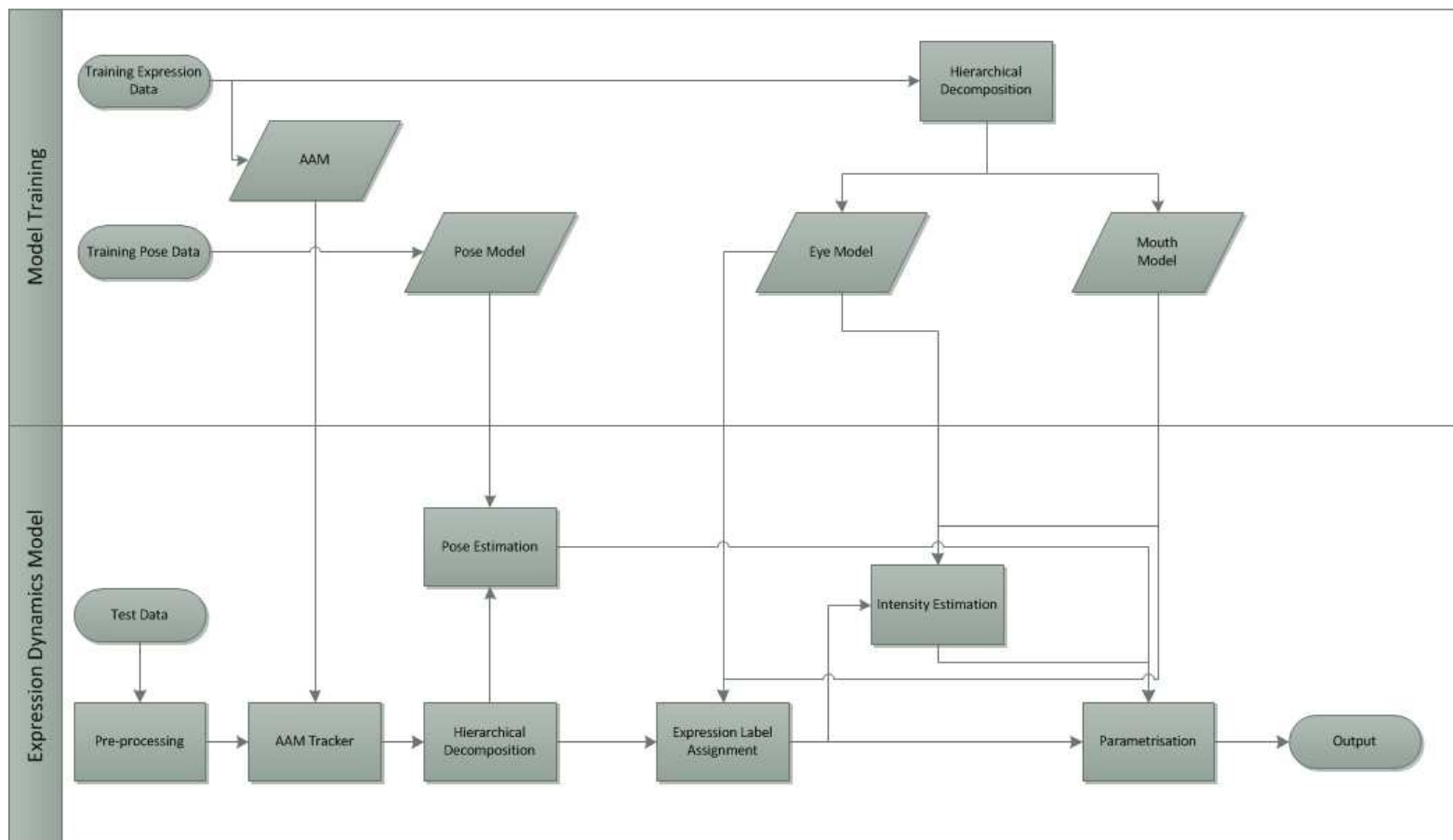


Figure 5.1: General overview of the process in our system.

5.1.2 Rule-based fusion

Rule-based classifiers are frequently used in machine learning because of the ease with which rules might be interpreted by humans. If the number of rules is relatively small, and the detection accuracy is sufficiently high, such classifiers are an optimal choice, because the reasons for their decisions can easily be verified [32]. Furthermore they do not require training, which in the case of large datasets can be time consuming. For optimal results, rules should be mutually exclusive, and their coverage should be exhaustive over the problem domain. This can be tricky for complex problems that cannot be easily decomposed into simpler, atomic units. Pantic and Rothkrantz [89] used a rule based classifier to determine the resulting AUs from a parametric description of the features. Our concern is at a higher level of abstraction, and its intent to combine, or fuse available information from the eye and mouth components. Given shape input feature vectors \mathbf{t}_{mouth} , \mathbf{t}_{eyeL} , \mathbf{t}_{eyeR} obtained from hierarchical decomposition described in Chapter 3, and resulting class-conditional probabilities for eye and mouth models defined by Equations (3.40) and (3.41), we define the discrete outputs for the mouth (M), left eye (EL) and right eye (ER) as follows:

$$\begin{aligned}
 M &= i, \quad i \in \{1^{(neutral)}, 2^{(smile)}, 3^{(grin)}, 4^{(open)}, 5^{(anger)}, 6^{(sad)}, 7^{(disgust)}\} \\
 ER &= j, \quad j \in \{1^{(squint)}, 2^{(neutral)}, 3^{(open)}\} \\
 EL &= k, \quad k \in \{1^{(squint)}, 2^{(neutral)}, 3^{(open)}\}
 \end{aligned} \tag{5.1}$$

where i , j and k correspond to the class, or label with the highest value. The final label assignment is performed by combining the information obtained from all three discrete outputs together. The final

expression label F is defined as:

$$F = \begin{cases} \textit{smile} & M = 2 \\ \textit{grin} & M = 3 \\ \textit{fear/surp} & (M = 4 \wedge (ER = 3 \vee EL = 3)) \\ & \vee (M = 4 \wedge ER = 2 \wedge EL = 2) \\ \textit{anger} & (M = 5 \wedge (ER = 1 \vee EL = 1)) \\ & \vee (M = 1 \wedge ER = 1 \wedge EL = 1) \\ \textit{sad} & M = 6 \wedge (ER = 1 \vee EL = 1) \\ \textit{disgust} & M = 7 \wedge (ER = 1 \vee EL = 1) \\ \textit{neutral} & \textit{otherwise} \end{cases} \quad (5.2)$$

where \vee represents logical OR and \wedge represents logical AND.

Note, that for some expressions, certain facial regions are not utilised. In the design of this classifier we have taken into consideration the findings of Nusseck et al. [81] regarding necessity and sufficiency of similar facial regions with respect to expression recognition. Their psychophysical experiments show that some expressions depend solely on a single facial region, where as for others dependence extends to the combination of such regions. For example for happiness and surprise the mouth region is sufficient, but for surprise both eyes and mouth are required. On the other hand, Pelachaud and Poggi [91] suggest that for surprise both mouth and eye regions are necessary. We adopted the latter line of thought, as firstly we consider that surprise can be represented by raising of the eyebrows only, and secondly the use of both regions provides some level of redundancy. The latter reason was driven mainly by practicality, because inaccurate data from the mouth region alone cannot provide reliable labels.

5.1.3 Severity Criterion

We have found [123] that severity, or intensity combined with an appropriate expression label, summarises the underlying expression dynamics. The majority of approaches are only interested in expression classification, and their parameterisation will only deliver a discrete on/off output. Un-

fortunately such a representation does not yield the required continuous response and fine-grained detail for the expressions in question for many applications.

To address this issue, our severity information metric is defined in terms of the combination of severities of the hierarchical subcomponents. For each of the eye and mouth components the information is measured in terms of corresponding low level, or intrinsic, behaviours. Given the high level final expression label F obtained during the fusion process in Equation (5.2), for each of the components, severity is defined by the marginal cumulative distribution function given by Equation (4.2) of the probability density model for the low level behaviour belonging to the j -th hierarchical component, where $j \in \{mouth, eye_L, eye_R\}$.

For the combined intensity value, if the high level label $F \in \{smile, grin\}$ then the severity S is given by:

$$S = \Phi_{mouth}(\mathbf{t}_{mouth}) \quad (5.3)$$

where $\Phi(\mathbf{t}_{mouth})$ is the the cumulative distribution of the probability density function of the mouth hierarchical component, and $S \in [0, 1]$. For expressions other than smile or grin the severity is given by the weighted combination of all of the facial regions:

$$S = \sum_j w_j \Phi_j(\mathbf{t}_j) \quad (5.4)$$

where $\Phi(\mathbf{t}_j)$ is the cumulative distribution of the probability density function of the classified expression component $j \in \{mouth, eye_L, eye_R\}$, w_j are weights such that $\sum w_j = 1$, and $S \in [0, 1]$. The choice of weights was determined experimentally and was set to (0.4, 0.3, 0.3) for the *mouth*, *eye_L*, *eye_R* components respectively, and reflects the amount of influence each has towards the overall intensity. We also experimented with Mahalanobis Distance (MD) as an alternative measure for the calculation of severity which is given by:

$$D_M = \sqrt{(\mathbf{t} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{t} - \boldsymbol{\mu})} \quad (5.5)$$

Unfortunately we have found that it is not sufficient for the task. Figure 5.2 shows two images that demonstrate gradual change for two expressions grin (top row) and fear (bottom row). We can see that our SC produces correct values compared to MD. This is caused by a lack of symmetry in PCA

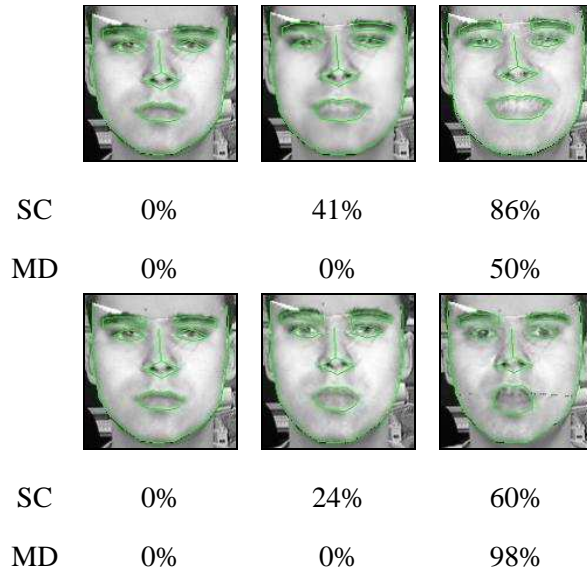


Figure 5.2: Selected frames from two expressions demonstrating gradual change for grin (top row) and fear (bottom row) and the severities associated with them using our method (SC) and naive approach based on (MD).

space and the fact that variation in each of the principal components is a combination of many factors, not just those caused solely by the expressions.

5.2 3D Animation Model

The output from our expression parameterisation system is used to operate a 3D animation model. We employ a blend shape, or morph based approach, where the resulting expression is created through the combination from the collection of existing sample bases [39]. Every character is required to have a set of predefined bases corresponding to the expressions we wish to model. Then any expression E is given by:

$$E = \sum_i \mathbf{w}_i \Gamma_i \quad (5.6)$$

where \mathbf{w} defines a weight vector and $\mathbf{w}_i \in [0, 1]$ for every i . Γ defines a set of morph bases corresponding to predefined expression states. Figure 5.3 shows an example of these morph bases for two different avatars (top and bottom). Although somewhat limited in the sense of available free-



Figure 5.3: An example of morph bases (left to right) for neutral, smile, grin, surprise/fear, anger for two different avatars (top and bottom).

dom, and requiring pre-rigged characters, such an approach offers several advantages. Firstly the representation is compact and independent of the animation engine, giving us the ability to model human and non-human characters alike. Secondly the complexity of the model and the number of parameters is relatively small compared to physics-based models, such as the one described by [90]. Reduced complexity and the small number of control parameters opens up possibilities for real-time animation. We demonstrate how our 3D animation model performs in the next section.

5.3 Experiment

Here we use our expression parameterisation approach together with the 3D animation model described in the previous section. These form part of our system outlined in Figure 5.1. For face tracking we employed the Cootes and Taylor [24] Active Appearance Model (AAM) tracker, which was trained using a set consisting of 1300 images and shapes (each with 74 landmarks), which con-

sisted of six basic expressions (neutral, smile, grin, sadness, surprise/fear, anger) and variations in pose. The resulting tracker was based on a person specific AAM for robustness under sparse training samples, and also for providing a better basis for capturing intricate expressions.

To test our system we used two test sequences **T1** and **T2** totalling 2850 frames altogether. In both of the sequences the exhibited expressions were interleaved with casual speech intervals. Although we did not specifically consider modelling speech, we felt its inclusion might provide a more realistic scenario, where expressions would naturally be separated by speech fragments. To evaluate the performance we compared our rule-based method with that of a holistic approach adopted by Huang and Huang [53] and Liu et al. [71], and with the Bayesian Network (BN) of Chuang et al. [19]. Although in the latter the BN was defined in the context of features and by design assumed independence between them, we will utilise the hierarchical models and assume independence between those. All of the methods were trained using the same training set consisting of 813 hand labelled samples, representing continuous changes of various facial expressions. For the BN we have performed supervised training. Figure 5.4 shows the label assignment results for test sequence **T1** and Figure 5.5 shows the label assignment results for test sequence **T2**. In both cases our rule based method provided the best results. All three methods were able to reliably assign labels to grin and surprise expressions, and this is partly due to the fact that those expressions were the most reliably and accurately tracked. However the holistic approach completely failed to recognise smile and anger expressions, whilst the BN method completely failed to recognise smile, and had more limited success generally. Unfortunately all of the methods completely failed to recognise sadness. These results highlight the ability of the rule driven selection in our hierarchical approach to provide the required level of redundancy and the ability to cope with noisy data. On the other hand, the holistic approach is not able to provide such redundancy and hence is unable to adapt, mainly due to its global nature and the resultant constraints. Similarly, the performance of the BN, although better than the holistic method, can be attributed to the training process, in which only correct data was used. Figure 5.6 depicts overall label assignment scores, and Table 5.1 summarises the results of the experiment. Due to the lack of standardised evaluation tests for continuous and gradual parameterisation the evaluation

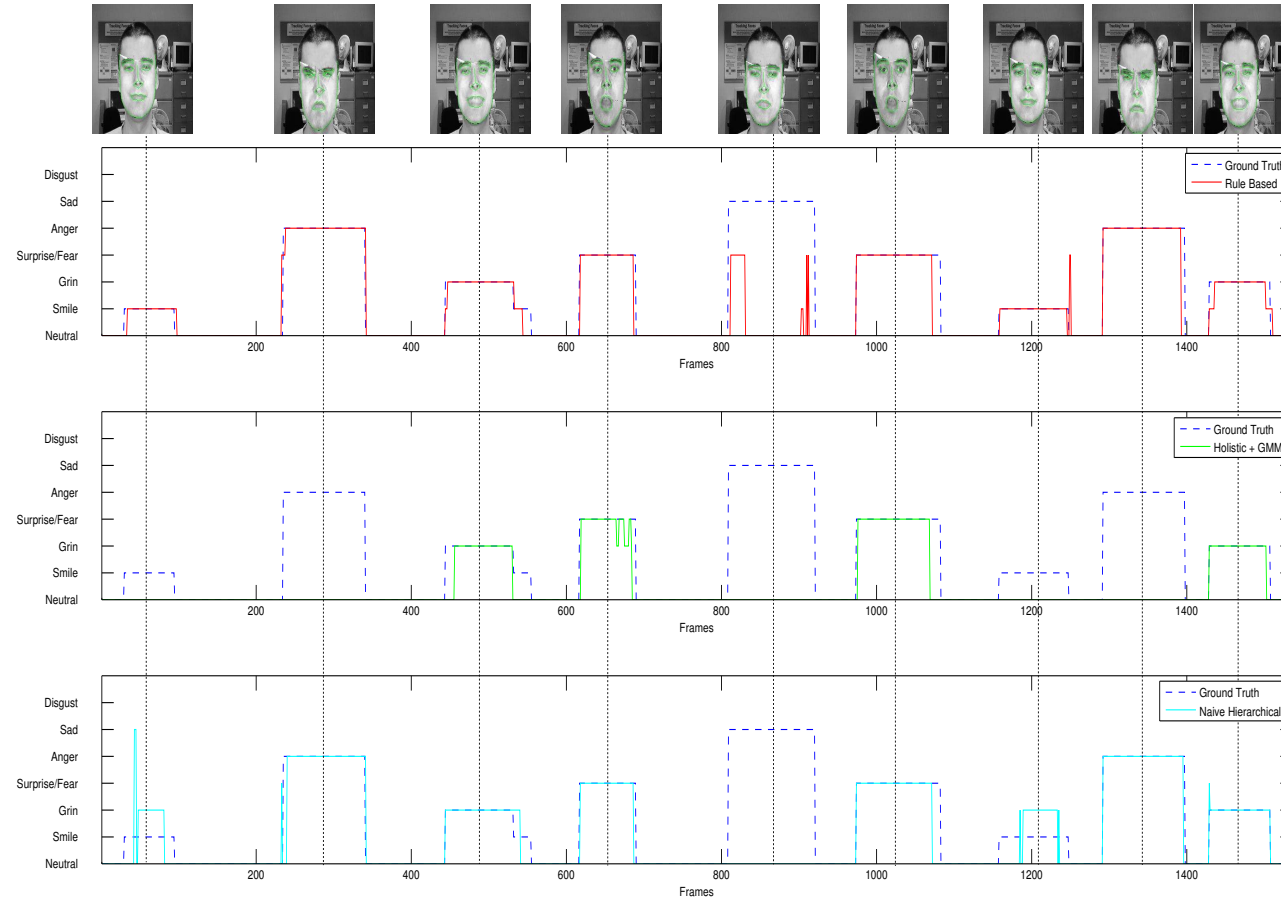


Figure 5.4: Label assignment together with the corresponding ground truth and selected keyframes for test sequence **T1** where the top row corresponds to rule-based, middle to holistic and bottom to BN approaches.

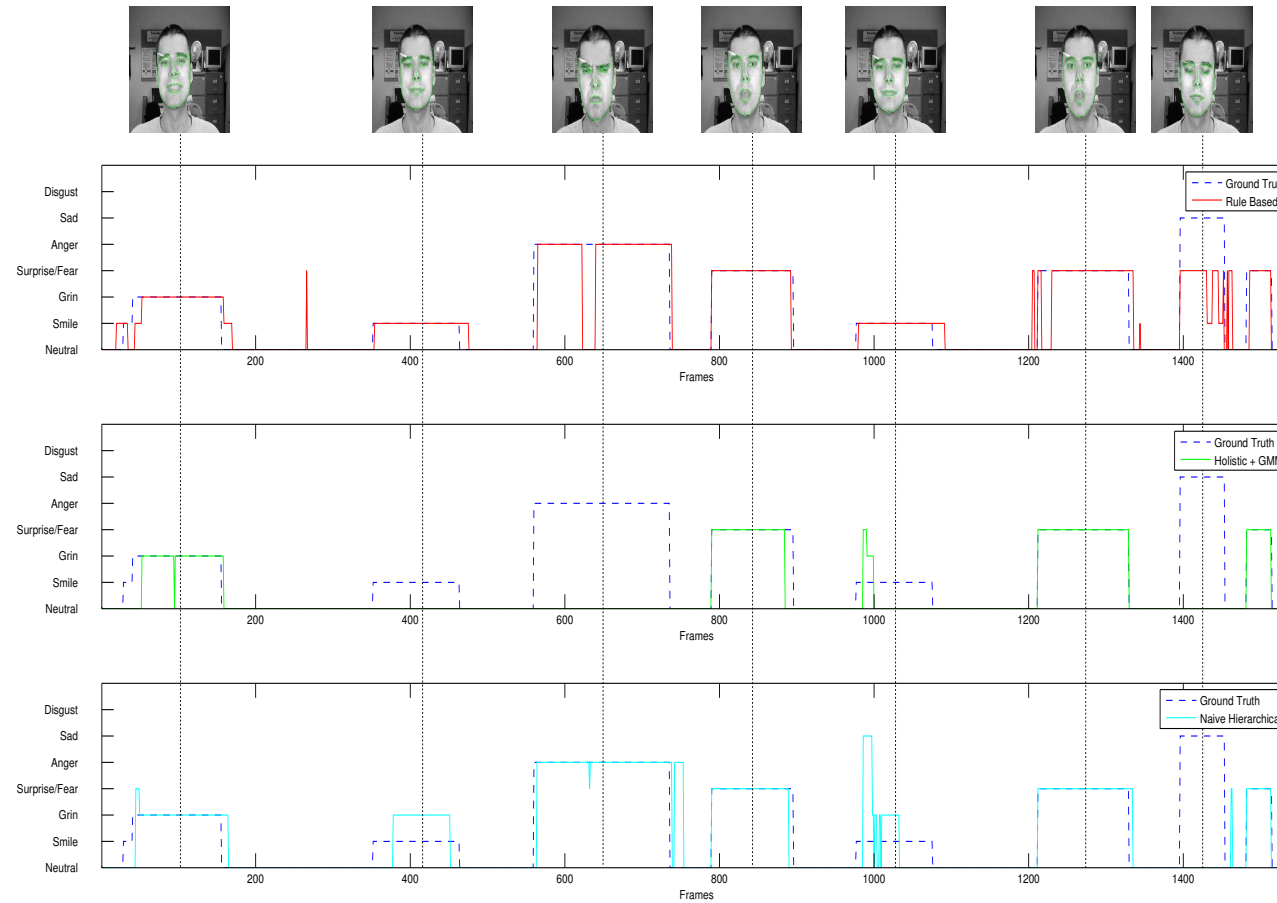
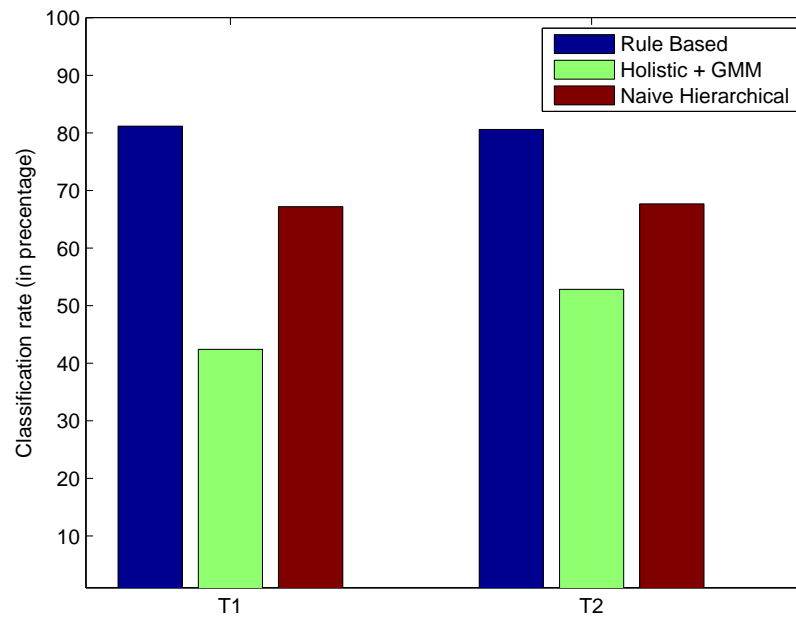


Figure 5.5: Label assignment together with the corresponding ground truth and selected keyframes for test sequence **T2** where the top row corresponds to rule-based, middle to holistic and bottom to BN approaches.

	Rule-based	Holistic ([53, 71])	Naive Hierarchical
seq T1	81.15%	42.40%	67.19%
seq T2	80.57%	52.82%	67.66%

Table 5.1: Table comparison of label assignments methods.

Figure 5.6: Final label assignment scores for test sequences **T1** and **T2**.

was based only on a simple label assignment scheme (expressions present/absent) as the compared approaches did not define the severity of the expressions.

Given the resultant labels from the best performing method, we next move on to estimating the expression intensity. The choice of the marginal CDFs for each of the hierarchical regions was determined by analysis of the underlying models, and choosing the one that contributes the most to the underlying intrinsic functionality rate of change. Similarly to the pose estimation technique,

we have smoothed the animation curves by convolving them with a Gaussian kernel. Figure 5.7 shows the resulting severities for each of the expressions for test sequence **T1** (bottom row) together with corresponding label assignment scores for rule-based approach (top row). Figure 5.8 shows the resulting severities for each of the expressions for test sequence **T2** (bottom row) together with corresponding label assignment scores for rule-based approach (top row).

Finally we assembled the complete parametric form and applied it to a virtual head avatar. Figure 5.9 shows selected frames from a casual conversation sequence. Within each of the boxes the left image corresponds to the currently tracked image frame with the AAM appearance mask superimposed on it. The image on the right corresponds to the synthetic avatar animated according to the classified expression. These results clearly show a useful correlation between input expressions and the avatar output.

5.4 Discussion

In this chapter we have presented an alternative way of describing expressions as a combination of intrinsic functionalities of three hierarchical component areas corresponding to the face, both eyes and the mouth. To formulate the final expression labels, we investigated the concept of fusion of information obtained from different hierarchical components. Rule-based classification offers an efficient and more robust alternative to more complex classifiers due to its simplicity and absence of training stage. We have shown that our approach performs better than the holistic approach of [53, 71] and the Bayesian Network (BN) approach of [19] where the relationships amongst the components were learned through a training process. This highlights the fact that holistic methods model the global nature of variations and cannot adapt to localised changes caused by inaccuracies in the data. Although BN performed better than the holistic approach, since it was based on hierarchical components, its ability to determine the correct labels was limited by the training set, which only contained correct samples.

Given the resulting labels, we have also explored intensity estimation of the corresponding expressions as a weighted combination of intensities of component intrinsic functionalities. This of-

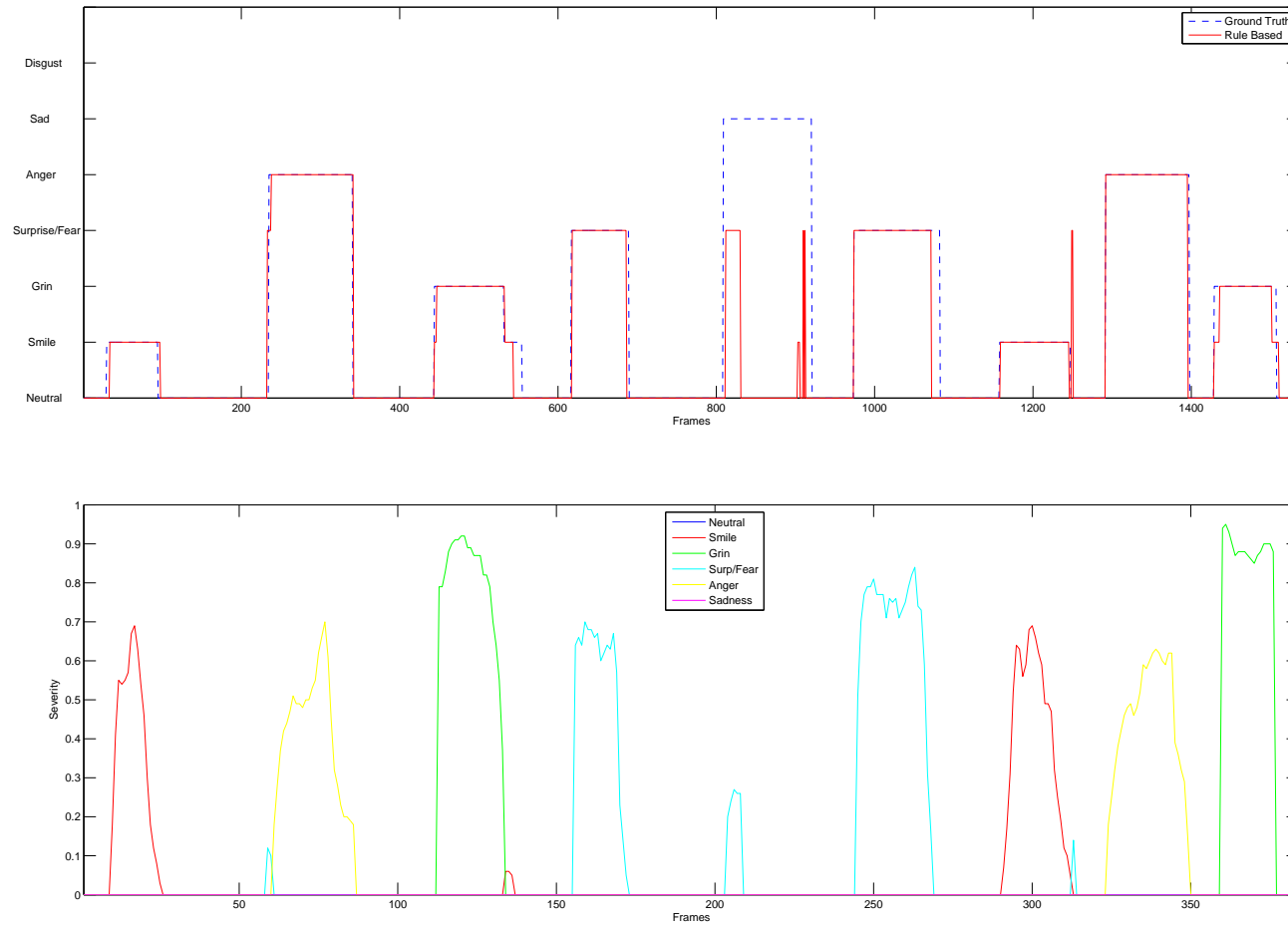


Figure 5.7: Severity curves for test sequence **T1** (bottom row) together with corresponding label assignment scores for rule-based approach (top row).

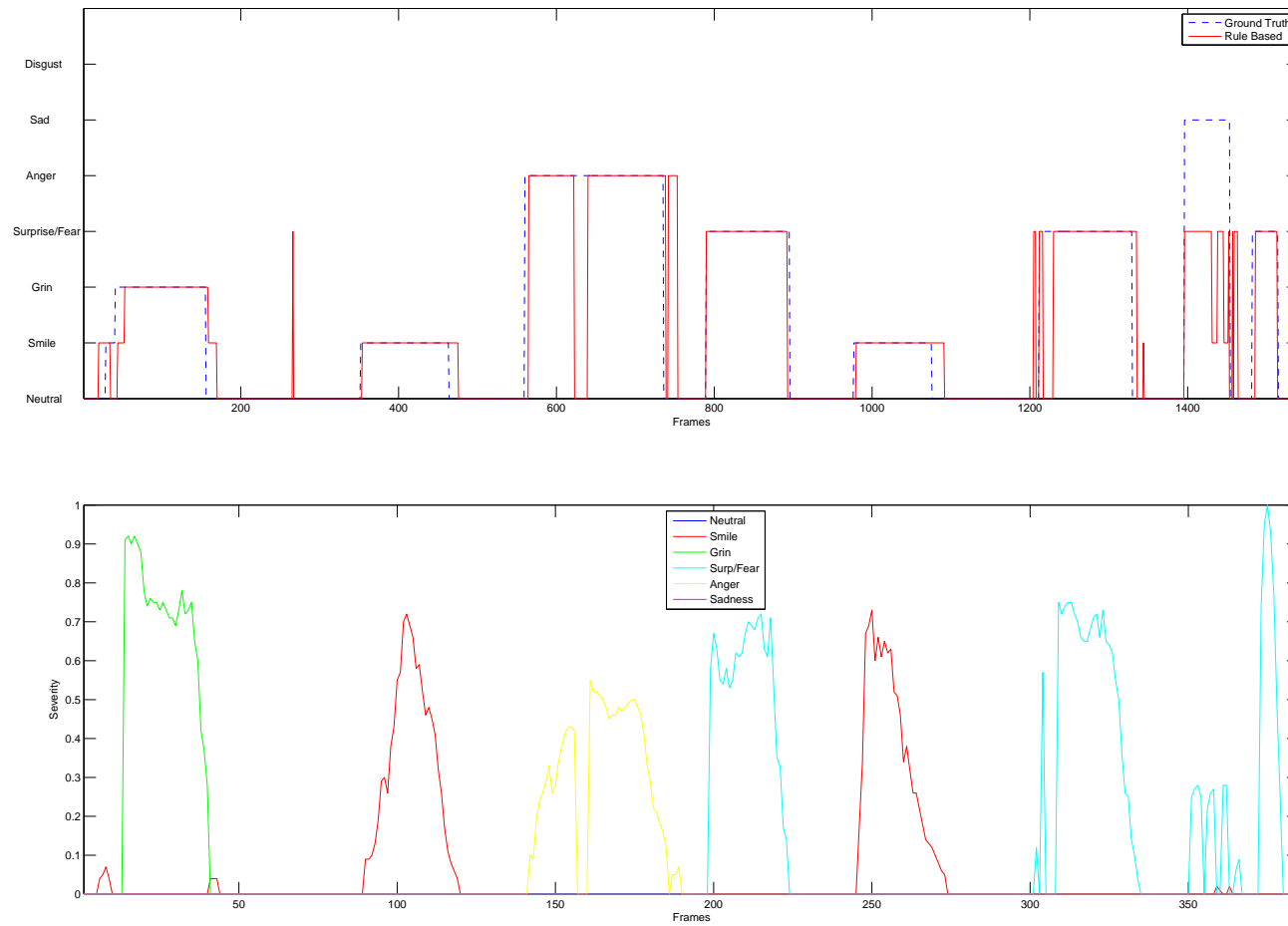


Figure 5.8: Severity curves for test sequence **T2** (bottom row) together with corresponding label assignment scores for rule-based approach (top row).

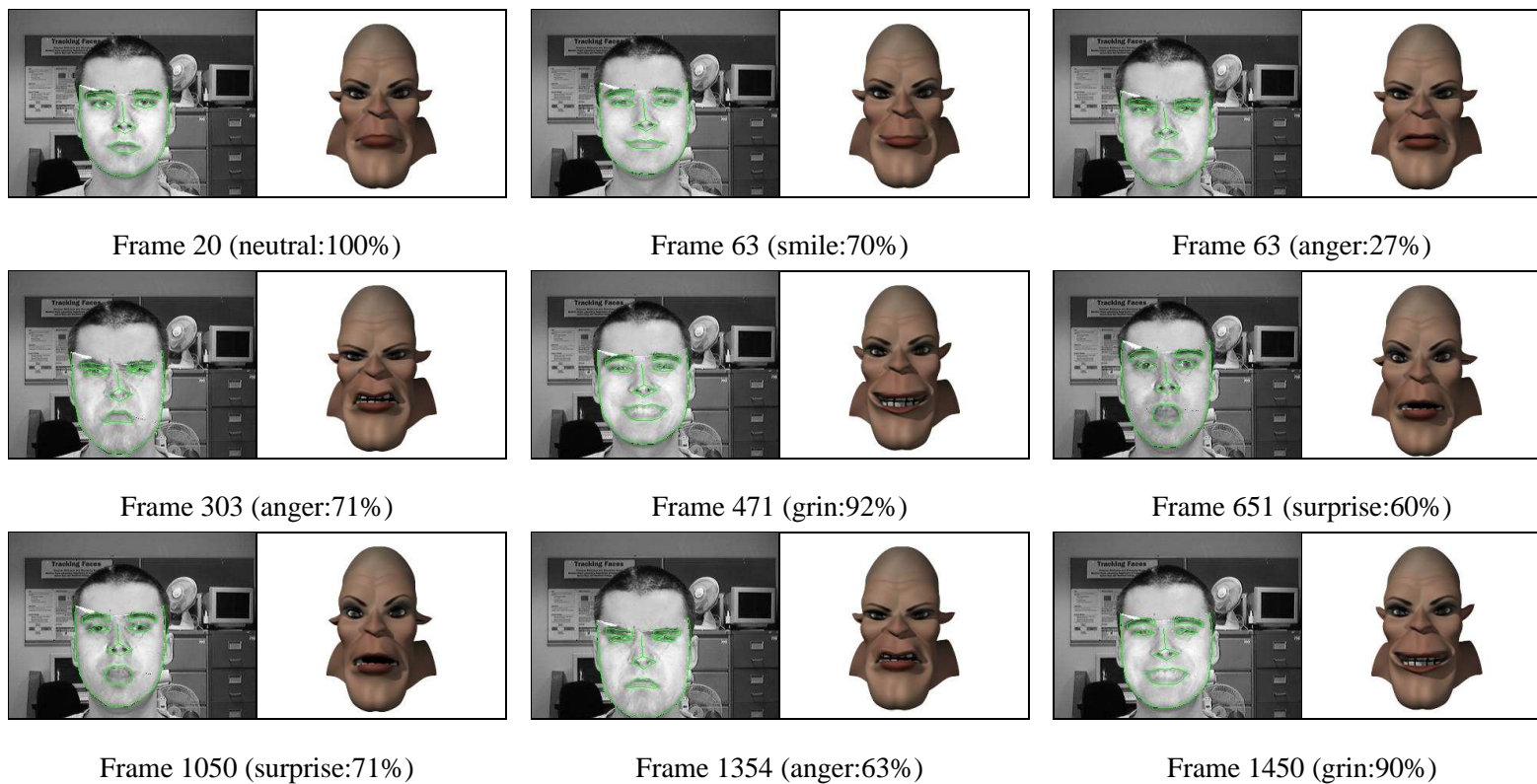


Figure 5.9: Selected frames from experiments on expression classification and avatar animation with corresponding labels and severity (as a percentage). Each of the images shows the tracked frame with AAM mask superimposed on it (left), and corresponding synthesised avatar (right).

ferred the ability to account for contributions of different facial regions towards the final value, and allowed us to measure the intensity of expressions, where only single regions trigger the activation.

Overall in this chapter we have demonstrated the possibilities of processing facial representations from input to output with evident success. The range of possible applications for such a system must surely be considerable.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The human face provides a most powerful medium for relaying emotions between humans and plays an important, if not crucial, role in our day to day social interaction. Due to the rapid advances of computer hardware and the growing popularity of personal computers in our everyday lives, machine analysis of facial expressions has been emerging as an active field for the last decade with growing interest in application fields such as human-computer interaction, computer animation, computer gaming and social networking.

In this thesis we studied methods for building a framework capable of modelling and parameterising facial expression dynamics. In particular three challenging issues have been addressed:

Hierarchical Feature Representation

We investigated facial feature representation based on the Active Appearance Model (AAM), and focused on its shape component also known as Point Distribution Model (PDM). Although shape alone cannot capture skin changes, it provides significantly lower feature and hence model dimensionality. This approach also provides invariance to illumination changes and some invariance to intra-person

variations. We have employed hierarchical decomposition of the holistic shape representation into subcomponents corresponding to the eyes and mouth; which we found to be the most salient facial regions. The underlying dynamics of each of those regions was represented by a Hierarchical Latent Variable Model (HLVM). We have found that both eye components can be sufficiently represented by a unified, single model further reducing the overall model dimensionality. In contrast to the holistic representation, our approach allows us to reduce the number of non-linearities present due to intra-feature correlations, and provide a more precise representation with respect to the intrinsic functionalities of the components, rather than final expression labels. In real world scenarios, where data will contain inaccuracies and noise, our approach provides better performance compared to a PCA based one.

Pose Estimation

Pose is an important, but somehow overlooked, part of expression dynamics. As our head rarely stays still, it is intrinsically connected with the way we perceive facial expressions. Some of the expressions, such as nods or shakes are even represented solely by the pose. Large head motion causes non-linearities in the shape space which makes PCA not sufficient for representing this subspace. Our proposed method integrates into our existing hierarchical framework, and employs a HLVM to model its underlying distribution, and is entirely based on $2D$ information obtained from a sparse, discretely sampled training set. As expressions can have significant impact on pose we have demonstrated that by making an appropriate choice of features we can reduce this impact. We have demonstrated how the underlying probability model can be used to estimate the pose and can generalise to estimate continuous pose from unseen samples. Such a model can also be used in the synthesis of arbitrary expressions at arbitrary viewpoints, where it serves as a morph, or warp basis. Finally prior knowledge of the pose information can also assist in an AAM fitting process, where due to self-occlusion parts of the available appearance information become unavailable.

Expression Dynamics Modelling

We have introduced a framework that is capable of extracting facial dynamics information and producing a parameterised version thereof. We investigated the fusion of information obtained from hierarchical components using rule-based and Bayesian Network (BN) classifiers. Based on the assigned labels and using an underlying hierarchical statistical model we estimate the intensity of the final expression as a combination of intensities of the intrinsic functionalities of the modelled facial regions. We show that our method performs better than the holistic approach. Finally we create a parametric description of the expression dynamics and apply it to animate a synthetic head avatar.

6.2 Future Work

So far we have discussed issues related to modelling of facial dynamics, and although progress has been made our work still has some limitations. Here we list those limitations and possible directions for future work:

- The AAM based tracking is not feasible in real world environments, due to the laborious labelling process. We would like to investigate alternative methods, which provide a more stable and reliable basis and ultimately help to reduce model preparation and training time.
- Another important component of facial dynamics is eye blinking. Combined with subtle head movements it defines what we learn to perceive as expressions from a very young age. Its absence makes the synthesised expressions unrealistic.
- Although our test sequences included fragments of speech, we have not focused on them specifically. In real world scenarios, it might not be possible to extract the visemes¹, due to the fast and usually blurred lip movements. However knowledge of the speech and non-speech segments would give us the ability to simulate, or synthesize lip movements in the synthetic counterpart hence enhancing the visual experience even more.

¹viseme is a representational unit used to classify speech sounds in the visual domain. It describes the particular facial and oral positions and movements that occur alongside the voicing of phonemes. (from wikipedia)

- Throughout the course of this work we have focused on a pre-defined universal set of expressions. We would like to expand this set to include a wider range of expressions including those caused by the head movements alone.
- We have focused on static expression parameterisation and its sufficiency for the task at hand. We do realise, that for some of the expressions, dynamic knowledge of how the face changes over time is important and enhances the amount of available information.
- So far we have shown that shape alone can provide a sufficient basis for estimating the dynamics of the expressions. In addition, we would like to analyse the appearance information, which would increase the amount of available information, allowing us to investigate FACS parametric descriptors.

Bibliography

- [1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, Prague, 2004.
- [2] M. Amin, N. Afzulpurkar, M. Dailey, V. Esichaikul, and D. Batanov. Fuzzy-c-mean determines the principle component pairs to estimate the degree of emotion from facial expressions. In *Fuzzy Systems and Knowledge Discovery*, volume 3613 of *Lecture Notes in Computer Science*, pages 481–481. Springer Berlin / Heidelberg, 2005.
- [3] K. Anderson and P. W. McOwan. A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics*, 36(1):96–105, 2006.
- [4] I. Bacivarov and P.M. Corcoran. Facial expression modeling using component aam models - gaming applications. In *International IEEE Consumer Electronics Society Games Innovation Conference*, pages 1–16, London, 2009.
- [5] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop*, volume 5, page 53, Madison, 2003.
- [6] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 223–230, Southampton, 2006.
- [7] M.S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods

- for fully automatic recognition of facial expressions and facial actions. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 592–597, Hague, 2004.
- [8] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. In *International Conference on Computer Vision*, pages 323–328, Bombay, 1998.
- [9] J.N. Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059, 1979.
- [10] F. Berisha, A. Johnston, and P. McOwan. Identifying regions that carry the best information about global facial configurations. *Journal of Vision*, 10(11), 2010.
- [11] M. Beszedes and P. Culverhouse. Comparison of human and automatic facial emotions and emotion intensity levels recognition. In *International Symposium on Image and Signal Processing and Analysis*, pages 429–434, Istanbul, 2007.
- [12] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *British Machine Vision Conference*, volume 2, pages 797–806, Leeds, 2002.
- [13] C. M. Bishop and M. E. Tipping. A hierarchical latent variable model for data visualisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):281–293, 1998.
- [14] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.
- [15] Y. Chang, C. Hu, and M. Turk. Manifold of facial expression. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 28–35, Nice, 2003.
- [16] Y. Chang, C. Hu, and M. Turk. Probabilistic expression analysis on manifolds. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 520–527, Washington DC, 2004.

- [17] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.
- [18] K. Cho, Y. Kim, and Y. Lee. Real-time expression recognition system using active appearance model and efm. In *Computational Intelligence and Security*, volume 4456 of *Lecture Notes in Computer Science*, pages 1078–1084. Springer Berlin / Heidelberg, 2007.
- [19] E. S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *Conference on Computer Graphics and Applications*, pages 68–76, Beijing, 2002.
- [20] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences. In *International conference on Multimedia and Expo*, volume 2, pages 121–124, Lausanne, 2002.
- [21] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 2003.
- [22] J. F. Cohn. Foundations of human computing: facial expression and emotion. In *International conference on Multimodal interfaces*, pages 233–238, New York, 2006.
- [23] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 227–232, Grenoble, 2000.
- [24] T. F. Cootes and C. J. Taylor. Statistical models of apperance for computer vision. Technical report, University of Manchester, Manchester, UK, March 2004.
- [25] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active apperance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [26] D. Cosker, S. Roy, P. Rosin, and D. Marshall. Re-mapping animation parameters between

- multiple types of facial model. In *International conference on Computer vision/computer graphics collaboration techniques*, pages 365–376, Rocquencourt, 2007.
- [27] D. W. Cunningham, M. Kleiner, C. Wallraven, and H. H. Bühlhoff. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception*, 2(3):251–269, 2005.
- [28] G. Donato, M. S. Barlet, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.
- [29] F. Dornaika and J. Ahlberg. Efficient active appearance model for real-time head and facial feature tracking. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 173–180, Nice, 2003.
- [30] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition using a particle filter. In *International Conference on Computer Vision*, volume 2, pages 1733–1738, Beijing, 2005.
- [31] Y. Du, W. Bi, T. Wang, Y. Zhang, and H. Ai. Distributing expressional faces in 2-d emotional space. In *International conference on Image and video retrieval*, pages 395–400, Amsterdam, 2007.
- [32] W. Duch, N. Jankowski, K. Grabczewski, and R. Adamczak. Optimization and interpretation of rule-based classifiers. In *Intelligent Information Systems IX*, pages 1–14, Bystra, 2000.
- [33] G. Edwards, T. Cootes, and C. Taylor. Advances in active appearance models. In *International Conference on Computer Vision*, volume 1, pages 137–142, Kerkyra, 1999.
- [34] P. Eisert and B. Girod. Model-based estimation of facial expression parameters from image sequences. In *International Conference on Image Processing*, volume 2, pages 418–421, Santa Barbara, 1997.

- [35] P. Ekman. *The argument and evidence about universals in facial expressions of emotion*, pages 143–164. Chichester: Wiley, 1989.
- [36] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.
- [37] P. Ekman and W. V. Friesen. Emfacs. unpublished manuscript, 1982.
- [38] P. Ekman, E. Rosenberg, and J. Hager. Facial action coding system affect interpretation dictionary (facsaidd) <http://face-and-emotion.com/dataface/facsaidd/description.jsp>, 1998.
- [39] N. Ersotelos and F. Dong. Building highly realistic facial modeling and animation: a survey. *Visual Computing*, 24(1):13–30, 2008.
- [40] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.
- [41] B. Fasel and J. Luetttin. Automatic facial expression analysis: a survey. *Journal of the Pattern Recognition Society*, 36:259–275, 2003.
- [42] B. Fleming and D. Dobbs. *Animating Facial Features & Expressions*. Charles River Media, 1998.
- [43] S. Gong, A. Psarrou, and S. Romdhani. Corresponding dynamic appearances. *Image and Vision Computing*, 20(4):307–318, 2002.
- [44] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991.
- [45] R. T. Griesser, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff. Psychophysical investigation of facial expressions using computer animated faces. In *Symposium on Applied perception in graphics and visualization*, pages 11–18, Tubingen, 2007.

- [46] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. In *British Machine Vision Conference*, pages 457–466, Kingston, 2004.
- [47] H. Gu and Q. Ji. Facial event classification with task oriented dynamic bayesian network. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 870–875, Washington DC, 2004.
- [48] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4:223–233, 2000.
- [49] A. Heap and D. Hogg. Improving specificity in PDMs using a hierarchical approach. In *British Machine Vision Conference*, volume 1, pages 80–89, Essex, 1997.
- [50] E. Hjelmas and B. Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–274, 2001.
- [51] T. Hong, , Y. Lee, Y. Kim, and H. Kim. Facial expression recognition using active appearance model. In *Advances in Neural Networks - ISNN 2006*, volume 3972 of *Lecture Notes in Computer Science*, pages 69–76. Springer Berlin / Heidelberg, 2006.
- [52] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3d head orientation from a monocular image squence. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 242–247, Vermont, 1996.
- [53] C. L. Huang and Y. M. Huang. Facial expression recognition using model-based feature extraction and action parameters classification. *Journal of Visual Communication and Image Representation*, 8(3):278 – 290, 1997.
- [54] O. Ikeda. Segmentation of faces in video footage using hsv color for face detection and image retrieval. In *IEEE International Conference on Image Processing*, volume 3, pages 913–916, Barcelona, 2003.

- [55] Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE International Conference on Systems, Man and Cybernetics*, 36(5):862–875, 2006.
- [56] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, Grenoble, 2000.
- [57] M. Khademi, M. T. M. Shalmani, M. H. Kiapour, and A. A. Kiaei. Recognizing combinations of facial action units with different intensity using a mixture of hidden markov models and neural network. In *International Workshop on Multiple Classifier Systems*, volume 5997, pages 304–313, Cairo, 2010.
- [58] S. Kimura and M. Yachida. Facial expression recognition and its degree estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 295–300, San Juan, 1997.
- [59] Kwang-Eun Ko and Kwee-Bo Sim. Development of advanced active appearance model for facial emotion recognition. In *IEEE International Symposium on Industrial Electronics*, pages 1019–1022, Seoul, 2009.
- [60] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 4, pages 3732–3737, Orlando, 1997.
- [61] S. Koelstra and M. Pantic. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, Amsterdam, 2008.
- [62] W. J. Krzanowski. *Principles of Multivariate Analysis*. Oxford University Press, 1988.
- [63] S. Kshirsagar, M. Escher, G. Sannier, and N. Magnenat-Thalmann. Multimodal animation

- system based on the mpeg-4 standard. In *International Conference on Multimedia Modelling*, pages 5–21, Ottawa, 1999.
- [64] D.T. Lee and B. J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Parallel Programming*, 9:219–242, 1980.
- [65] K.K. Lee and Y. Xu. Real-time estimation of facial expression intensity. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 2567–2572, Taipei, 2003.
- [66] J. J. Lien. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1998.
- [67] J. J. Lien, T. Kanade, J. Cohn, and C. Li. Subtly different facial expression recognition and expression intensity estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 853 – 859, Santa Barbara, 1998.
- [68] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. In *Computer Vision and Pattern Recognition Workshop*, page 80, Washington DC, 2004.
- [69] G. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Joint Symposium on Neural Computation*, San Diego, 2006.
- [70] Gwen Littlewort, Ian Fasel, Marian Stewart Bartlett, and Javier R. Movellan. Fully automatic coding of basic expressions from video. Technical report, Univeristy of California, San Diego, INC MPLab, 2002.
- [71] W. Liu, J. Lu, Z. Wang, and H. Song. An expression space model for facial expression analysis. In *Congress on Image and Signal Processing*, volume 2, pages 680–684, Washington DC, 2008.

- [72] Z. Liu and B. Guo. *Face Synthesis*, chapter 12. Springer-Verlag New York, Secaucus, NJ, USA, 2005.
- [73] S. Lucey, A. B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. In *Face Recognition Book*. Pro Literatur Verlag, Mammendorf, Germany, 2007.
- [74] M. Luthi, T. Albrecht, and T. Vetter. Probabilistic modeling and visualization of the flexibility in morphable models. In *Mathematics of Surfaces XIII*, volume 5654 of *Lecture Notes in Computer Science*, pages 251–264. Springer Berlin / Heidelberg, 2009.
- [75] J. Ma, J. Yan, and R. Cole. Cu animate: Tools for enabling conversions with animated characters. In *International Conference on Spoken Language Processing*, pages 197–200, Denver, 2002.
- [76] G. Maestri. *[digital] Character Animation 2*. New Riders, 1999.
- [77] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, November 2004.
- [78] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):780–788, 2002.
- [79] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *Automatic Systems for the Identification and Inspection of Humans*, volume 2277, pages 12–21, San Diego, 1994.
- [80] J. Ng and S. Gong. Composite support vector machines for detection of faces across views and pose estimation. *Image and Vision Computing*, 20(5-6):359–368, 2002.
- [81] M. Nusseck, D. W., Cunningham, C. Wallraven, and H. H. Bühlhoff. The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision*, 8(8):1–23, 2008.

- [82] J. Ostermann. Animation of synthetic faces in mpeg-4. *IEEE Computer Animation*, pages 49–55, 1998.
- [83] T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using hmm. In *IEEE International Conference on Automatic Face and Gesture Recognition*, page 442, Washington DC, 1998.
- [84] C. Padgett and G. Cottrell. Representing face images for emotion classification. In *Advances in Neural Information Processing Systems*, volume 9, pages 894–900, Denver, 1996.
- [85] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, 2007.
- [86] M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics*, 36(2):433–449, 2006.
- [87] M. Pantic and J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
- [88] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. *Image and Vision Computing*, 18(11):881–905, 2000.
- [89] M. Pantic and L.J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(3):1449–1461, 2004.
- [90] F. Parke. *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Salt Lake City, Utah, December 1974. UTEC-CSc75-047.
- [91] C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 13(5):301–312, 2002.

- [92] S. M. Platt and N. I. Badler. Animating facial expressions. In *Conference on Computer graphics and interactive techniques*, pages 245–252, Dallas, 1981.
- [93] Yogesh Raja. *Adaptive Visual Sampling*. PhD thesis, Queen Mary, University of London, 2010.
- [94] J. Reilly, J. Ghent, and J. McDonald. Non-linear approaches for the classification of facial expressions at varying degrees of intensity. In *International Conference on Machine Vision and Image Processing*, pages 125–132, Maynooth, 2007.
- [95] J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.
- [96] N. Sarris, N. Grammalidis, and M.G. Strintzis. Fap extraction using three-dimensional motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(10):865 – 876, 2002.
- [97] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. In *International Conference on Computer Vision*, pages 221–230, Beijing, 2005.
- [98] C. Shan, S. Gong, and P. W. McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing*, volume 2, pages 914–917, Genoa, 2005.
- [99] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.
- [100] Caifeng Shan. *Inferring Facial and Body Language*. PhD thesis, Queen Mary, University of London, 2007.
- [101] M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *International Joint Conference on Pattern Recognition*, pages 408–410, 1978.

- [102] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.
- [103] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.
- [104] Y. Tian, T. Kanade, and J. F. Cohn. Recognising action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):1–19, 2001.
- [105] Y. Tian, L. Brown, A. Hampapur, S. Pankanti, A. Senior, and R. Bolle. Real world real-time automatic recognition of facial expressions. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2003.
- [106] Y. Tian, T. Kanade, and J. F. Cohn. *Facial Expression Analysis*, chapter 11. Springer-Verlag New York, Inc., Secaucus, NJ, 2005.
- [107] M. E. Tipping and C. M. Bishop. Mixture of probabilistic component analysers. Technical report, Dept of Computer Science and Applied Mathematics Aston University, Birmingham B4 7ET, UK, 1998.
- [108] Y. Tong, W. Liao, and Q. Ji. Inferring facial action units with causal relations. In *Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1623–1630, New York, 2006.
- [109] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- [110] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Computer Vision and Pattern Recognition Workshop*, page 149, New York, 2006.
- [111] P. Vanger, R. Honlinger, and H. Haykin. Applications of synergetic in decoding facial expressions of emotions. In *International Workshop on Automatic Face and Gesture Recognition*, pages 24–29, Zurich, 1995.

- [112] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1692–1698, Hawaii, 2005.
- [113] A. Watt. *3D Computer Graphics*. Addison Wesley, 1999.
- [114] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. In *Conference on Computer Vision and Pattern Recognition*, pages 535–542, Washington DC, 2004.
- [115] Jing Xiao, T. Kanade, and J.F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 156–162, Washington DC, 2002.
- [116] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6): 636–642, 1996.
- [117] M. Yand, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 34–58, 2002.
- [118] P. Yang, Q. Liu, and D.N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *International Conference on Computer Vision*, pages 1018–1025, Kyoto, 2009.
- [119] M. Yeasin, B. Bullot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508, 2006.
- [120] L. Zalewski and S. Gong. A probabilistic hierarchical framework for active appearance models. In *Symposium on Language Speech and Gesture for Expressive Characters, AISB Convention*, pages 12–19, Leeds, 2004.

- [121] L Zalewski and S. Gong. Synthesis and recognition of facial expressions in virtual 3d views. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 493–498, Seoul, 2004.
- [122] L Zalewski and S. Gong. A statistical virtual head animator. In *IEE International Conference on Visual Information Engineering*, pages 31–38, Glasgow, 2005.
- [123] L Zalewski and S. Gong. 2d statistical models of facial expressions for realistic 3d avatar animation. In *Conference on Computer Vision and Pattern Recognition*, pages 217–222, San Diego, 2005.
- [124] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
- [125] C. P. Zhang and F. S. Cohen. Component-based active appearance models for face modelling. In *International Conference on Advances in Biometrics*, pages 206–212, Hong Kong, 2006.
- [126] Y. Zhang and Q. Ji. Facial expression understanding in image sequences using dynamic and active visual information fusion. In *International Conference on Computer Vision*, volume 2, page 1297, Washington DC, 2003.
- [127] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.
- [128] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.
- [129] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE*

International Conference on Automatic Face and Gesture Recognition, pages 454–459, Nara, 1998.

- [130] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju. Real-time automatic deceit detection from involuntary facial expressions. In *Conference on Computer Vision and Pattern Recognition*, pages 1–6, Minneapolis, 2007.
- [131] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. Technical report, University of Maryland, 2001.
- [132] Z. Zhu and Q. Ji. Robust real-time face pose and facial expression recovery. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 681–688, New York, 2006.